NTNU

Norwegian University of
Science and Technology

Disambiguation of named entities using a novel gamified framework

Brikend Rama

Applied Computer Science
Submission date: June 2017
Supervisor: Mariusz Nowostawski, IDI

Norwegian University of Science and Technology
Department of Computer Science

Preface

This research represents the “Disambiguation of named Entities using a novel gamified framework”, the basis of which consolidates a gamified system established on empirical research and standard implementation practices. It represents additional research work in improving the field of semantic web and advancing natural language processing techniques.

The project was undertaken as a master thesis within the Department of Computer Science (IDI) at the Norwegian University for Science and Technology (NTNU). The targeted audience of this research study includes gamification enthusiast, semantic web and natural language processing practitioners who pursue new methodologies and approaches for improving the respective aforementioned fields.

01-06-2017

Acknowledgment

My sincerest gratitude go to my supervisor, Mariusz Nowostawski, for his extensive support, academic and expert advice and continuous feedback throughout the process of conducting this difficult research study. I would also like to thank Rune Hjelsvold and Simon McCollum for their academic feedback and advice which truly helped in improving the quality of this research.

Finally, I must express my very profound gratitude to my family and friends for their unconditional moral support and continuous encouragement throughout the process of writing this thesis. Thank you.

B.R.

Abstract

The content generated on the web originates from diverse sources with the main purpose of serving updated information to the Internet user. Every piece of information generated is valuable and must be easily traced by modern search engines. Semantic meta-data as a mechanism for providing meaning to the generated content is the de-facto requirement for improving search accuracy and facilitating information discovery on the web. This research represents an attempt for advancing the field of semantic web in terms of providing an approach for generating semantic information to the substantial number of unstructured documents available on the web. The main objective is to utilize the potential of human computation as a source for improving the performance of supervised and semi-supervised algorithms in the respective field. Performance improvements are achieved through the generation of large-scale qualitative annotation data. Being a time and resource intensive process to be carried out by expert annotators, ordinary non-expert annotators must be encouraged for contribution.

Gamification as an increasingly popular approach for leveraging human computational power has been investigated in this research study. It represents the ultimate tool for encouraging human annotators for contribution in exchange for an engaging and attractive game. The disambiguation of recognized named entities within the content of unstructured web documents represents the problem elaborated in this work. Therefore, the implementation of a generic and scalable gamified named entity disambiguation framework demonstrating the capabilities of non-expert users in generating large-scale annotation data represents the main qualities composing this research study. We specifically focus on benefiting from gamification as a powerful and prominent approach for leveraging human computation. Significant and confident results acquired through experimental user studies support the idea that gamification can successfully leverage human computation for collaboratively solving complex problems. This comes as a result of game design which is based on on empirical research, psychological theories for motivation and standard practice of implementation.

Contents

List of Figures

List of Tables

Abbreviations

AI Artificial Intelligence

AMQP Advanced Message Queuing Protocol

CET Cognitive Evaluation Theory

CRC Collaborative Resource Creation

GWAP Games With A Purpose

KB Knowledge Base

LOD Linked Open Data

NED Named Entity Disambiguation

NER Named Entity Recognizer

NLP Natural Language Processing

NP Nondeterministic Polynomial Time

OMCS Open Mind Common Sense

RDF Resource Description Framework

REST Representational State Transfer

SDT Self-Determination Theory

WSD Word Sense Disambiguation

1 Introduction

1.1 Topic

Natural Language Processing (NLP) tasks generally fall into the category of artificial intelligence and machine learning problems and are considered to be relatively complex [4, 5]. During the last decades, researchers have been focusing on solving problems ranging from named entity disambiguation, language translation, word sense disambiguation and anaphora resolution [6, 7, 5, 8, 4]. This research study is particularly focused on the problem of named entity disambiguation. Tackling this problem results in improving, among others, searching accuracy on the WEB which contains a large proportion of unstructured documents that lack semantic encoded meta-data. Incorporating such information within Web documents is crucial for improving the performance of search engines [9].

The incredible rapid advancements in information processing technologies and task automation using intelligent machine learning algorithms has opened up the discovery of novel "solutions" to high complexity problems for which machines were not able to solve before. However, despite these advancements, machine learning algorithms (also called supervised approaches) have been struggling on achieving high levels of quality and performance accuracy [4]. Such requirements are crucial and inevitable in fields like natural language processing, speech recognition, semantic web and the like.

The specific problem investigated in this research study is an interweaving problem within natural language processing and semantic web. Namely, we study the problem of Named Entity Disambiguation and the potential of using gamification in order to attract non-expert contributors for generating annotation data in a collaborative way. Named entity disambiguation, or NED for short, is the process of linking a real world object (i.e. a named entity) appearing in textual data with a knowledge base. A Knowledge Base (KB) is a machine-readable resource for the dissemination of information which is used to optimize information collection, organization and retrieval for an organization or the general public [10]. Some of the most frequently used knowledge bases are Dbpedia[1], FOAF[2], Google Knowledge Graph[3], Geonames[4] and many others generated during the past decade [11].

[1] Dbpedia Knowledge Base http://wiki.dbpedia.org/

[2] Friend Of A Friend Vocabulary http://xmlns.com/foaf/spec/

[3] Google Knowledg Graph API https://developers.google.com/knowledge-graph/

[4] Geonames Vocabulary http://www.geonames.org/

These knowledge bases are all part of the Linked Open Data Cloud initiative which aims at providing more complete answers to search engines as new data sources appear on the WEB [11]. An important and quite complex step for NED is to disambiguate a named entity by finding a candidate (among many) from the knowledge base that best describes the entity, based on the context in which it occurs. The concept of bridging documents on the WEB with knowledge bases is helpful for linking the large amount of raw and noisy data present on the WEB. This concept also contributes to Berners-Lee's proposed vision for the Semantic Web [5]. However, because of the ambiguous nature of entities, identifying and correctly linking entities with their corresponding counterparts on the knowledge base still remains a challenging task for machines [12].

To help assist and improve automatic algorithms in getting better at linking ambiguous entities, human input must be leveraged as a validation and quality assurance mechanism. In this context, assuring and maintaining high quality of annotations, users that take part in the validation process have to be either linguistic experts or trained annotators. However, very little research has been done towards supporting non-expert users in the process of creating semantically-enriched content [13]. A not so common approach to collaborative resource creation which is investigated in this research study, is intrinsically motivating users to create annotations by using a so-called Game With A Purpose (GWAP). Using a GWAP we are able to produce annotations as a byproduct of users playing the game [14]. Providing that the game is entertaining enough to attract sufficient players, according to Poesio et al. [15], it should be possible to carry out large-scale annotations of documents at smaller cost compared to other approaches such as crowdsourcing. This research study is primarily focused on investigating the potential of a gamified system for generating qualitative and accurate annotations by non-exert annotators. The gamified system is build on top of a complete NED framework and basis its design on theoretical and psychological models for user motivation and engagement. As such, we investigate on game design elements and techniques to intrinsically motivate users to do annotations without using any form of payment incentives, thus keeping cost at the lowest level possible. Finally, the generated annotations by the non-expert users (players) shall be used by artificial intelligence or machine learning algorithms as training data and also by other tools that aim at enriching unstructured web documents with semantic content.

1.2 Keywords

Gamification, GWAP, Named Entity Disambiguation, Entity Linking, Human Validation, Game Design, Semantic Web

1.3 Problem Description

This research study will try to answer two problems identified in the literature.

In Natural Language Processing (NLP), supervised approaches that use machine learning or other artificial intelligence techniques perform best compared to semi-supervised and unsupervised approaches [4]. Their performance highly depends on large amounts of annotated data (i.e. training data). This data is used by supervised approaches for either training the algorithm or evaluating the quality of annotations. The aforementioned training data is usually acquired by linguistic experts or trained annotators in a manual fashion. Consequently, the amount of training data required by a supervised algorithm to train its network is enormous which makes the gathering process very time- and cost-intensive, yet important and necessary. As a result, the process of creating large-scale annotated training data has yet remained a long-standing barrier for many areas of NLP [16].

Avoiding the idea of manually creating annotated corpus is not considered as a solution to the problem, since automatic approaches are still immature on performing such a complex task [4]. Therefore, to be able to find the answer to the first problem, it is necessary to investigate different manual techniques that will make the process of creating annotated data sets less tedious and cost intensive. Since it is being dealt with the human factor, additional problems arise in terms of motivation, training and data quality assurance. This inevitably leads this research to address these human factor problems and attempts to investigate on potential techniques on how to train and motivate non-expert users to perform large-scale and high-quality annotations. This research study has been conducted as a motivation for finding an optimal solution to the problems stated above.

1.4 Justification, motivation and benefits

The content on web pages, news articles, blog posts and other internet data consists of hundreds and thousand mentions of named entities such as people, organizations, locations and other relevant concepts. These entities are often ambiguous, meaning that the same name can have many different interpretations. Identifying these entities and linking them with a corresponding KB that is part of the Link Open Data (LOD) initiative, results in several benefits for information processing and retrieval systems such as search engines.

Despite the impressive enhancements of search engines in the last decade, information searching is still dependent on keyword-based searching. This searching technique usually does not fully meet the users needs due to insufficient content meaning on the web documents [17]. Since the techniques used by traditional search engines are based on straightforward matches of terms within unstructured text, understanding the context of the query created by the user is not taken into consideration [18]. As a result users get frustrated by having to adjust their query terms to retrieve the desired results. The proposed solution to this problem is semantic search [9]. Semantic search best operates when web documents contain semantic meta-data, which in turn allows the discovery of deeper meanings and relationships of specific query terms rather than relying on exact keyword matches [18]. Previous research suggests that attaching semantic meta-data to unstructured web documents clearly improves precision in search engines with maximum recall [18]. The problems addressed by our research study will contribute and have a significant impact towards improving semantic search.

Since the performance of supervised approaches for NED rely heavily on the availability of large training corpora, having a system that engages human users in an interactive and fun way can generate such corpora in a short period of time with minimal costs. According to Green et al. [19], knowledge captured by a specific annotated corpus is often not transferable to another task, even when it is the same NLP task but different language. This increases the importance of having a system which supports the generation of training data at minimal costs.

Furthermore, research suggests that turning a task into a GWAP has shown to increase quality of results and higher user engagements, thanks to the users being stimulated by the playful component [16]. Our motivation is to create a gamified framework model that is much in tune with the efforts of the Open Mind Initiative[5] [17]. Open Mind Initiative focuses on the collection of data from internet users and suggests using such data for training and improving machine learning algorithms.

Several other systems will benefit by having a reliable and accurate named entity disambiguation system. Information extraction, information retrieval, content analysis, question answering systems and knowledge base population are some of the applications where named entity disambiguation is considered as the initial step towards improving their accuracy and overall performance [20]. Potential stakeholders benefiting from the results of this research study include: gamification practitioners, semantic web enthusiasts, developers and researchers looking for alternative approaches for utilizing human computation within the fields of NLP and semantic web.

[5] Open Mind Initiative http://wiki.p2pfoundation.net/Open_Mind_Initiative

1.5 Research Questions & Hypothesis

Automatic named entity disambiguation has been extensively researched by previous studies, therefore, the focus of this research study is not directly improving the task itself. Instead, we focus on providing means that will help future researchers easily gather up-to-date training data for improving the accuracy of NED systems. With that being said, our main focus is finding a suitable approach for leveraging human input as a validation mechanisms for generating trustful and qualitative training data for supervised algorithms. The following research questions and hypothesis will help us conclude whether using gamification and non-expert annotators is a suitable approach for improving some of the supervised and semi-supervised algorithms in natural language processing:

- What are the underlying qualities of a NED framework for supporting the generation of trustful annotation data by ordinary non-expert annotators? How motivated are users in performing annotations using the non-gamified version of the framework?
- How much can proximity context features such as bigrams, neighbor entities and topic keywords contribute to informing human annotators for making correct disambiguation?
- What game mechanics can be employed in the entity disambiguation task so that high levels of engagement are achieved while still maintaining annotation quality? How do they affect player intrinsic motivation?

1.6 Contributions

First and foremost, this research study contributes to the research field by elaborating and unfolding the lessons learned from applying gamification on non-gaming contexts. The design decisions, utilized technology, techniques, assessment methodology and all the other aspects constituting this work represent another attempt to gamification which can potentially help others in the process of designing an efficient and highly useful GWAP. As such, our primary contribution is providing a model which represents best practices on how to accurately apply gamification on non-gaming contexts. We design a GWAP that engages players with its well-designed, task-oriented game elements that contribute a great deal to player intrinsic motivation and generation of qualitative annotations. With a GWAP designed and implemented on top of a microservice framework, we open up doors for further contributions to the research field. Small modifications or additional integration to the microservice framework, it will be possible to generate annotation data for other problems in the filed of NLP such as (potentially) language translation and speech recognition. Using this system, researchers and developers will be able to generate training data at minimal costs with data quality comparable to linguistic experts or trained annotators.

As a result of applying gamification on a complex systems such as named entity disambiguation, this research study provides additional contributions by offering a generic microservice architectural framework as a tool for disambiguating named entities. More specifically, the framework deals with extracting named entities, automatically linking them with knowledge bases and providing techniques for formulating the surrounding context of an entity. All this generated information is presented to the human annotators for accurately disambiguating ambiguous entities with the right KB candidate. In summary, the implemented framework represents a huge part of this work and is a fundamental process necessary to be undertaken in order to get to our first and main contribution.

The complete implementation of the Named Entity Disambiguation Framework is open and available for everyone who is interested in utilizing it for similar or other research problems[6]. Additionally, the complete gameplay of Fastype (our gamified version of named entity disambiguation) is demonstrated in a video which is uploaded on Youtube and can be accessed through the following link[7].

1.7 Outline of Chapters

In order to understand how named entity disambiguation is performed as well as understanding the underlying theoretical models and approaches used to build such framework, a theoretical background explaining notions and concepts with regards to NED, context definition and gamification is necessary. Chapter 2 provides the theoretical background on which the work of this thesis is build upon. Chapter 3 goes into a detailed explanation of the named entity disambiguation framework excluding the GWAP which is discussed later on in Chapter 4. Both these chapters provide detailed elaboration on the methodology, related work, setting up and conducting experiments and also analysis on the results obtained. Discussions on the limitations, strengths, weaknesses and implications of this study are elaborated on Chapter 5. Chapter 6 concludes our work and looks upon potential future work to further advance the field.

[6] AnnotateMe Framework https://github.com/brikendr/AnnotateMeFramework

[7] Fastype Gameplay https://www.youtube.com/watch?v=FWJkHvHfjOU

2 Background

The findings acquired by this research study are based on empirical research and principles of natural language processing, semantic web and game theory. Having a clear understanding of the concept of gamification, the different components that compose the framework, conceptualizing the term *context* within the scope of our problem are the main points discussed in this chapter. Additionally, since the main contribution of this work is mainly focused on the gamification of non-game related tasks, detailed insights on game design models and theories will be introduced in this chapter accordingly.

2.1 Games With A Purpose (GWAP)

Gamification represents the idea of using game design principles for transforming a system which was originally not created as a fun activity into an engaging and interactive game with a purpose [14, 21]. In this study we elaborate on the potential of applying game theory and design principles into the task of named entity disambiguation. In doing so, we are able to achieve large-scale annotation data that can be used either as training data for supervised NLP algorithms or as a tool for enriching web documents with semantic meta-data.

The proposed gamification approach of this research study and the field in which this technique is being applied, in literature is referred to as Games With A Purpose (GWAP) [14]. GWAP is one of the many approaches of gamification. The concept of gamification, as seen by researchers, involves applying elements of *gamefulness, gameful interaction and gameful design* with a specific intention in mind. According to Seaborn et al. [21], *gamefulness* refers to the lived experience, *gameful interaction* refers to the object, tools and contexts that bring the feeling of gamefulness while *gameful design* refers to the practice of creating a gameful experience. On the other hand, the aim of GWAP is to entertain players while they complete tasks that the system does not know, for most of the part, the correct answer. Usually, a well designed GWAP harvests the knowledge of their players and acquires the solution to the underlying problems as a byproduct of players playing and interacting with the game [14]. A major challenge in gamification is providing appropriate feedback at appropriate times during the gameplay. Overcoming this challenge results in making players feel engaged whilst the game uses their knowledge and experience to find solutions to the underlying problem. Understanding the motivation of players in this scenario is key to the success of a GWAP [22].

Being able to guarantee, to some extent, that players will be engaged and motivated to play the game, certain psychological needs have to be fulfilled so that players have the feeling of being immersively away from the real world and fully concentrated on the game. The answer to this question lies on theoretical cognitive theories such as the widely used Self-Determination Theory (SDT) and its respective sub-theories [23]. The process of designing the game for this particular research problem has been completely based on theoretical foundations and psychological theories for motivation (i.e. SDT) in order to reach a state where players experience the feelings of being entertained.

SDT, is a macro theory of human motivation that is essentially concerned with the potential for social contexts to provide satisfying experience. In SDT, the importance of competence (i.e. outcome control), autonomy (i.e. agency) and relatedness (i.e. connecting with others) are emphasized as the main factors to intrinsic motivation. Intrinsic motivation denotes the pursuit of an action because it is inherently enjoyable or interesting. In contrast, extrinsic motivation is defined as doing something due to a separable outcome, such as pressure or *extrinsic rewards* in the form of payment incentives or verbal feedback (ex. praise). Competence on the other hand, signifies the perceived extent of ones own actions as a cause of desired consequences and, as a psychological factor, is increased when the corresponding action is met with direct and positive feedback. It must be noted here that feelings of competence will not increase intrinsic motivation unless accompanied with a sense of autonomy. To affect feelings of autonomy, people must experience their actions and behaviour as self-determined rather than controlled by the system or an outside source. In support towards this concept, Cognitive Evaluation Theory (CET), a sub-theory of SDT, suggests that activities/actions foster greater intrinsic motivation when they provide goal-oriented tasks and an effort-full challenge. [22, 2, 23]

2.2 Linked Open Data (LOD) and the Web of Data concept

Disambiguation of named entities by linking them with a knowledge base has many benefits that contribute in bringing the idea of Semantic Web of Data and the Linked Open Data principles closer to realization. Berners-Lee et al. [11] define LOD as follows:

> "Linked Data is simply about using the Web to created typed links between data from different sources. It refers to data published on the Web in such a way that it is machine-readable, its meaning is explicitly defined, it is linked to other external data sets and can in turn be linked to and from other data sets.". [11]

Furthermore, Tim Berners-Lee also argues that "*The first step to semantic web is* *putting data on the web in a form that is understandable by machines or converting it to that form*" [11]. This is an important concept to shift our focus to, because it provides the means of publishing data on the web in such a way that all data published in the LOD will become part of a single global data space. The concept of Semantic Web should not be understood in alignment with the old and traditional *web of pages* where the main concern is putting data on the web. Semantic Web is about making links which should encourage exploration of the semantically connected web of data not only by humans but also by machines. Since the current WEB consists of large amount of unstructured information, based on the principles of LOD and Web of Data, it is necessary to convert this information to the desired form so that the goal of having a Semantic Web of Data is finally reached. "*While semantic web, or web of data, is the goal for the end-result of this process, Linked Open Data provides the means to reach that goal*" [11].

Our work contributes to the idea of semantic web and LOD principles in a way that it provides a gamified system that efficiently generates annotations used for enriching unstructured web documents with semantic meta-data. The generated annotations refer to named entities associated with a link to the corresponding KB candidate that describes the meaning of the entity (in our case Dbpedia). Bauer et al. [24] called Dbpedia "the semantic sister" of the most popular online encyclopedia in the world: Wikipedia. This makes Dbpedia one of the largest cross-domain knowledge bases extracted from the English edition of Wikipedia [24]. During the time of writing, Dbpedia is said to be the nucleus of the LOD cloud. It is one of the few KB that has most in-links and out-links to other KB published on the LOD cloud [10, 24]. The so-called LOD-Cloud, covers more than an estimated 50 billion facts from many different domains like geography, multimedia, biology, politics, academia, energy and the like. Datasets published in the cloud are described with a unique language called "Resource Description Framework" or RDF for short. It is a widely adopted standard for describing metadata as well as providing the means of structuring and linking data that describe things in the real world [11]. Figure 1 represents the LOD Diagram as of 2017 and is constantly updated and maintained by the Linked Open Data initiative community [1].

2.3 Named Entity Disambiguation (NED)

The Named Entity Disambiguation term refers to the process of identifying potential entity mentions in textual data and linking them with the corresponding candidate from a KB. The disambiguation part from the complete term refers to selecting an entity candidate which accurately represents the named entities' meaning based on the surrounding context. The so called concept of *Wikification* as explained by Trani et al. [12] is a similar approach to NED except that in their case the link

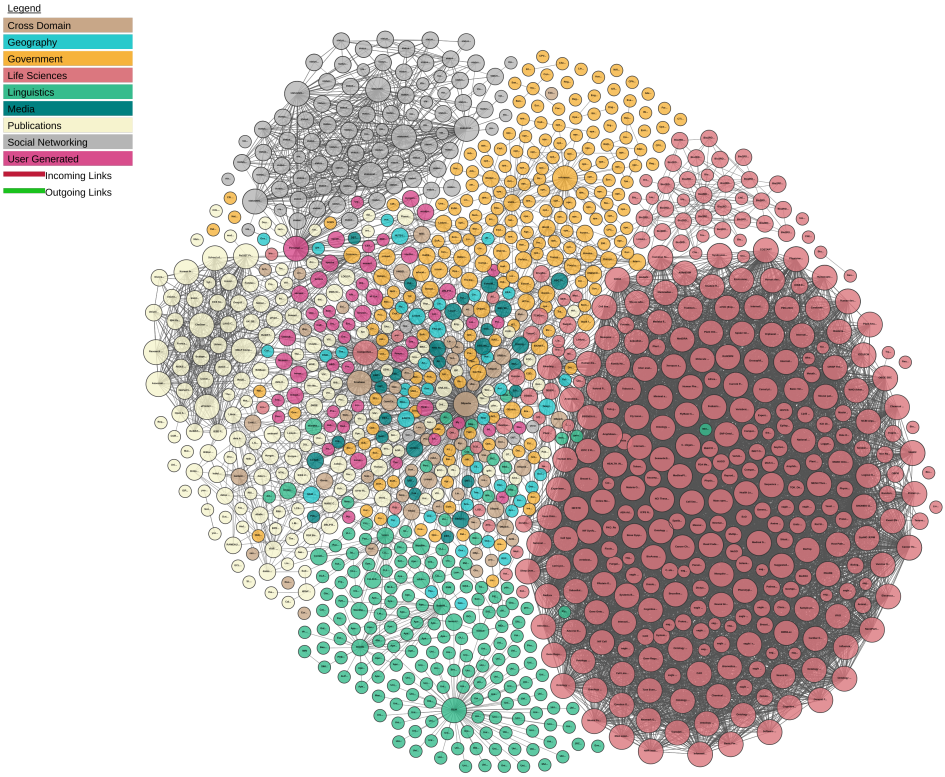


Figure 1: Linked Open Data cloud diagram 2017 [1]

associated with the entity mention corresponds to a Wikipedia page instead of an actual KB link.

Furthermore, NED can also be explained by analyzing another similar NLP problem, that is, WSD which stands for Word Sense Disambiguation. WSD represents the task of determining the correct meaning (sense) of a word in a given context [25]. WSD is similar to entity linking/disambiguation[1] in the sense that both problems refer to finding the correct reference of the spotted mention in an unstructured text document [25].

Defining the correct sense of a word or entity mention means knowing how to formulate the surrounding context. The surrounding context gives critical information to either the human or machine annotator for making an informed decision regarding disambiguation. The aforementioned process is considered as an AI-complete problem for machines, which in analogy to NP-completeness in complexity theory is a problem whose difficulty is equivalent to solving central problems of Artificial Intelligence (AI) [4]. When attacking these problems in an automatic fashion using AI algorithms, prior knowledge is required. According to Navigli [4], given a set of words, the procedure followed by a WSD system starts by applying techniques which make use of one or more sources of knowledge to associate the most accurate senses with words in surrounding context. In analogy, NED and WSD can be seen as classification tasks where the candidate words or senses are the actual classes. Usually an automatic classification algorithm is applied to assign a class to each named entity or word occurrence. The association should come as a result of making a decision that is based on evidence from the surrounding context and from potential external knowledge sources such as dictionaries [4]. Although the approach investigated in our research study relies on manual human annotation, improving the automatic supervised classification techniques is the ultimate goal provided with enough training data.

Automatic approaches on the other hand can be classified in three different classes:

- Unsupervised,
- Semi-Supervised, and
- Supervised

Among these three categories, supervised approaches have proven to perform best in terms of accuracy and quality of annotations [26]. Unsupervised approaches usually rely on unlabeled corpora, and do not utilize manually sense-tagged data to provide a sense choice for a word in context [4]. Semi-supervised approaches,

[1] Please note that linking and disambiguation will be used interchangeably throughout the text, but will refer to the same conceptual idea of disambiguating an entity mention with a knowledge base

just like the name implies, also rely on unlabeled corpora. In addition to that, semi-supervised approaches use various classifiers which are trained on a smaller set of trained samples.

In contrast to the previous approaches, supervised methods use machine learning techniques to learn classifiers from labeled training sets. Furthermore, feature sets such as Part-of-Speech (POS) of neighboring words, local collocations[2], syntactic patterns and other global features are used as strong classification features for the supervised methods [25]. The fact that the later approach leads in terms of accuracy and performance does make it a favorable choice to use for discovering different solutions to NLP problems similar to NED and WSD. However, due to the data scarcity problem, relying on large-amount of training data for different domains, tasks and languages cannot be seen as a realistic assumption. Therefore, significant manual effort is required [4]. According to Sanderson [27], improvements in the performance of information retrieval systems would be observed only if problems such as NED and WSD would perform at a level of at least 90% accuracy [27, 4]. The only possible way of achieving high levels of performance on these kind of tasks is by training supervised algorithms with large amounts of training data. Investigating on techniques and methodologies that would assure the generation of large-scale training data while keeping costs at minimal levels and still maintaining quality is the focus of the upcoming chapters. Before proceeding to the next section, understanding what the term *context* refers to in our particular problem scope is the topic of the next subsection.

2.4 Defining Context

Extracting information about the surrounding context of a word, an entity or even the context of the document as a whole (topic context) is one of the most important and yet most difficult tasks to achieve in NLP. The surface form *Texas*, according to Wikipedia, can refer to no more than twenty different named entities that can potentially describe the surface form based on the context it occurs. It may refer to the University of Texas, a Texas British Pop-Band, the US State of Texas or a novel named Texas written by Jams Michner [28].

Context has the power of being virtually anything, and it can be seen as a container in which the phenomenon resides [29]. It represents the parts of a discourse that surround a word or passage and can clarify the meaning or the interrelated conditions in which something exists or occurs [30]. Bontas and Paslaru [30] argue that contexts does not represent actual situations but it represents the perspective of an agent of the situation, since context is considered to be a partial approximation of a complete state of the world given a time point [30, 29]. Many studies see

[2] Collocations are also known as bigrams which represent a meaningful combination of two words

contextual information as a path that leads supervised algorithm or human annotator to make a clear decision on the disambiguation process of ambiguous surface forms (words).

Furthermore, besides eliminating ambiguities, context may be used for completing the missing information in natural language utterances [30]. The level of impact that context has in the performance of NLP tasks, has taken the attention of many researchers who have been trying to formulate or define the context for many years [31, 32, 33, 26, 34, 35]. However, context processing largely depends on the application domain, and the procedures used to formulate it are way too specific to be used in a generic scope. Therefore, Bontas and Paslaru [30] state that no clear and common methodology exists (yet) for the development of context-aware applications irrespective of the domain they belong to.

Some of the most common used features for disambiguating the sense of an ambiguous word and also defining the surrounding context of a word include: surround words and their Part-Of-Speech (POS) tags, topic keywords, content bigrams and various syntactic properties [36]. Topic keywords are considered as topical features that represent the general context of a document in which the ambiguous word resides. Unlike local features, topical features define the general topic of the document and represent a more generic context [4]. On the other hand, local feature such as bi-grams and surround words are important to pin down more specific contextual information. Bigrams are ordered pairs of words that are judged statistically significant by a measure of association. They often provide very specific unambiguous clues regarding the content of a context [34]. Navigli [4] states that deciding on the appropriate size of context (the number of bigrams, surround words, topic keywords etc) is an important factor in the development of NLP tasks such as NED and WSD. Inappropriate formulation of the context size is known to negatively affect the disambiguation performance of these tasks [4].

Understanding the theoretical foundations and ideas explained in this chapter is crucial to understanding and reasoning the contributions provided by this research study. In the next chapter, the implemented named entity disambiguation framework, being the initial stage before proceeding with gamification, will be explored in great detail. We start of by introducing the reader with state-of-the-art in the respective field and proceed with technical and conceptual analysis of the different models and elements that build up the system.

3 AnnotateMe - An Entity Disambiguation Framework

The work carried out by this research study can be logically divided into two parts that are closely interconnected with each-other. The first part represents the ground work on top of which we experiment with different data and methodologies in order to get answers to our first two research questions. Specifically, the first part represents the implementation of a named entity disambiguation (NED) framework. Implementing the framework was crucial as it corresponds to the system which we try to gamify and transform into a GWAP. Having accurate representation of the identified entities, their corresponding knowledge base candidates and the surrounding context, compose the elements of the framework on which the success of the game design highly depends on. The first part of this research study is explained in great detail throughout this chapter by exploring previous related work specific to our problem. Furthermore, we continue by explaining the architecture of the framework, its underlying components, the methodology used to gather the data, prepare the experimental user study and analyze the results.

3.1 Background

The Named Entity Disambiguation process is usually composed of two different modules: named entity recognition and candidate generation which also does the entity linking and disambiguation. However, in this research study we extend the number of modules composing the framework in order to provide more information to the human annotators for conducting the disambiguation process. The additional modules implemented in our framework include:

- a module for extracting contextual clues around the named entities called *Context-Clue Generation*,
- a module for generating *topical clues* describing the general topic of the document in which the entity resides,
- a *Data Preparation* module which prepares the data for the human annotators,
- a module for generating candidate links from the knowledge base called *Candidate Generation* module, and finally
- the *AnnotateMe Interface* which in turn uses human annotators to validate the generated links

For a better understanding, lets take an example of a text fragment and explain what the different modules produce as an outcome.

> "While Apple is an electronics company, Mango is a clothing one and Orange is a communication one." - excerpt taken from KORE50 Dataset

The text fragment above consists of surface forms (named entities) with a very ambiguous nature that is genuinely hard for automatic annotators to disambiguate. The responsibility of the entity recognition module is to identify the corresponding named entities in the text fragment. These entities represent real-world objects and usually fall into one of the following categories: Organization, Location and People. Running the named entity recognition module on the above text fragment would result in the identification of the following entities: Apple, Mango and Orange (all three being identified as Organizations). During this step, a commonly known technique for identifying the entities is by using classifiers [37]. Our framework utilizes the Stanford Named Entity Recognizer (Stanford NER) for this purpose which is known as a CRFClassifier [38]. According to Finkel et al. [38] the recognizer uses a general implementation of a linear chain Conditional Random Field (CRF) sequence models. These models are trained using labeled data and are generally classified as supervised approaches. Our framework uses SanfordNER[1] with a standard CRFClassifier for the English language trained with features for 3 classes in particular (Organization, Location and People). However, the classifier can be extended for recognizing additional classes and in other languages as well, but this problem is out of the scope of this research study and therefore we use the standard classifier. It is important to note that StanfordNER is one of the tools used by the framework for recognizing entities within text fragments. The complete process will be explained later in Section 3.3.

After the named entities have been identified by the recognition module, the framework proceeds by extracting contextual clues for each individual entity. Document keywords are also extracted during this stage simultaneously. The importance of appropriately formulating the surrounding context in which the entity mentions occur has been explained in Section 2.4. Therefore, the Context-Clue Extraction and Topic-Keyword Extraction modules represent the crucial part of the information presented to the human annotator in order to make correct disambiguation. From the text fragment above, an accurate context clue for disambiguating the named entity *Apple* would be *electronics company*. From an annotation point of view, the extracted contextual clue such as *electronics company* for entity *Apple* would provide sufficient information for the human annotator to decide whether the entity *Apple* refers to the plant or Apple Inc. Topic keywords on the other hand, keywords that represent the general context of the document, are more useful

[1] Stanford NER Package `https://nlp.stanford.edu/software/CRF-NER.shtml`

when processing larger text chunks. In the text fragment above, the most appropriate and useful topic keywords would be: Apple, Orange, Mango and company.

The final step for completely resolving the text fragment example provided above consists of candidate generation and entity disambiguation. The former is done in an automatic fashion by utilizing an automatic annotator whereas the later is done by asking human annotators to pick the correct candidate for each entity. If we take the entity *Apple* as an example, the candidate generation module generates the following candidates:

- Apple (Plant, Species, Eukaryot)
- Apple Records (Company, RecordLabel)
- Apple Inc. (Organization, Company)
- Apple II (Computer)

A weighted score is correspondingly assigned to each candidate by the automatic annotator. This score represents the level of confidence which is a numerical value from 0 to 1. The candidate which has the greatest confidence score is the best representative for the entity Apple as judged by the automatic annotator. There are cases in which the automatic annotator fails to correctly disambiguate the entity by picking the wrong candidate, or not providing the right candidate in the list at all. It is the responsibility of a human annotator to decide on the correct candidate from the list (if listed) based on the contextual information provided as short clues.

The aim of this section was to establish a general understanding of the different modules composing the framework and their corresponding responsibilities on building the foundations for an effective and qualitative named entity disambiguation work-flow. Section 3.3 will analyze and explain the underlying technical details for each module.

3.2 Related Work

Named Entity Disambiguation is not a new problem and previous research have tried to improve it using various approaches. These approaches range from supervised automatic algorithms that rely on classifiers, machine learning algorithms to manual approaches that rely on human input for manually performing the disambiguation step. Therefore, in this section we try to summarize previous research studies with special emphasis in human-related approaches. Research work in automatic techniques are also discussed since our ultimate goal is contributing to the improvements of supervised approaches for NED until an acceptable level of accuracy is reached.

3.2.1 Automatic Approaches for Entity Disambiguation

One of the earliest attempts to model an Entity Disambiguation process (recognition and linking of surface forms) was performed by [7]. They modeled a system called Wikify! which, given an input document, was able to identify the important concepts in text and link them to the corresponding Wikipedia pages. They have utilized a sense inventory (Wordnet[2]) and a link probability algorithm for disambiguating the ambiguous surface forms and linking them with the correct wiki page. Wikify's approach for disambiguating a surface form is by extracting features from the phrase and its surrounding context and compares it with training examples extracted from the entire Wikipedia. Linden [8] links named entities with a KB by unifying Wikipedia and Wordnet. They use four Wikipedia sources for collecting information about the surface forms, namely: entity pages, redirect pages, disambiguation pages and hyperlinks in Wikipedia articles. This information is then used to generate candidate list for each entity mention. Linden does the disambiguation process by combining the following measures: link probability, semantic associative, semantic similarity and global coherence [8]. However, the drawback of these approach is that they require massive pre-processing effort (parsing the entire Wikipedia) [39].

Other research studies [40, 10], have used a controlled vocabulary for identifying entity mentions in the document. For the disambiguation process, Turian [40] used two sets of features, namely *link probability* of an entity mention and *commonness* of the candidates. In order to guarantee the accuracy of entity linking, the candidates selected during the disambiguation process have to be strongly related with the target entity (that means higher values of commonness and link probability). [40]

The EDL Framework [41] is a similar approach compared to Wikify! where the disambiguation process is done using a combination of search engine results

[2] Wordnet Lexical Database https://wordnet.princeton.edu/

and knowledge base repository mining. Their framework consists of three steps: querying a KB for identifying potential candidates for the entities extracted in text, querying search engines for the same purpose and finally comparing the results from these two steps and output the best matching candidate for a particular entity. This is an unsupervised approach that relies only on features for disambiguating candidate entities. [41]

Unlike previous approaches, Hoffart et al. [42] argue that the key for further improvements in the named entity disambiguation process it to jointly consider multiple mentions when ranking the candidates. They argue that, when disambiguating an entity, the framework should consider also other named entities in a collective manner in order to select the correct candidate describing the entity [42].

Alchemy API is a framework for semantic annotation developed by Watson LAB[3]. Alchemy API analyzes WEB or textual content by using built-in NLP techniques, machine learning algorithms and other complex linguistic, statistical and neural network algorithms. The Alchemy API framework provides functionalities similar to our framework except that no human validation is used here. We used Alchemy API for generating document topic keywords as part of our context-clue generation modules.

Dbpedia Spotlight [10] is a system for automatically annotating text documents with Dbpedia[4] URLs. The goal of the service is to provide comprehensive and flexible solutions for entity annotations by offering a cross-domain vocabulary that can describe entities with diverse nature. Similar to [40], they depend on a controlled vocabulary for recognizing entity mentions in text. In particular, they use the LingPipe Exact Dictionary-Based Chunker which is based on Hidden Markov Models [10]. Regarding the candidate selection process, they rely on their own localization dataset for determining candidate disambiguation for each entity mention. However this step does not decide on the correct candidate, it only filters out irrelevant options. The disambiguation step consists of a supervised approach using a vector representation of different context features around the surface form. They have used Vector Space Model (VSM) for modeling each DBpeida candidate as a multidimensional space of words represented as a vector. Dbpedia Spotlight[5] provides an open source and free to use Web Service API that allows third-party applications to run queries and retrieve annotations with links pointing to Dbpedia concepts. We use Dbpedia Spotlight as our utilized automatic annotator for generating entity candidates identified in textual content.

Collective disambiguation was also used by Chabchoub et al. [43]. They uti-

[3] IBM Watson Alchemy API https://www.ibm.com/watson/alchemy-api.html

[4] Dbpedia Knowledge Base http://wiki.dbpedia.org/

[5] Dbpedia Spotlight Demo http://demo.dbpedia-spotlight.org/

lize an open source NER system in combination with an open source automatic annotator for recognizing entities in text. Similar to our named entity recognition module, they develop matching and filtering algorithms for improving the recognition process in terms of precision and recall. Candidates for each entity mention are generated by querying Dbpedia Spotlight. For ambiguous entities, where more than one candidate is retrieved, the disambiguation is done by taking into account the other entity mentions that have been already disambiguated in the text. [43]

3.2.2 Human-Centered Approaches for Entity Disambiguation

Automatic techniques for NED, just like any other classification or learning algorithms, have not reached the level at which they can simulate the way humans think and make links between concepts and ideas [27, 4]. However, these techniques are being improved continuously by providing training data that the algorithms can learn from. Thus, getting closer to the ultimate goal where the human knowledge can be reproduced and taken advantage of. This is a strong justification why many research studies (referring to our study as well) are continuously trying to leverage human input until the automatic approaches are considered mature.

Asking human annotators for validating the links generated by automatic approaches has been the focus of many research studies. ELIANTO supports human labelling of semi-structured documents by asking users to annotate entity mentions and then ranking those entities based on the perceived relevance or salience [12]. Khan et al. [17] conducted a user study where they asked participants to re-validate the system's accuracy by tagging concepts identified by the semantic annotator (Alchemy API) with their corresponding meaning. Milne et al. [39] used crowdsourcing for validating the linking accuracy of their system that used machine learning for generating candidates. Van Veen et al. [44] implemented a system for named entity linking for dutch historical documents. They used machine learning and rule-based techniques for the linking process, whereas for the validation process they asked library employees to make correction and also add missing links. All of these systems rely on users' voluntary incentive for contributing annotations. However, the voluntary incentive of participants cannot be taken for granted and therefore, participation of users in such systems in a long-time period is questionable.

Loomp OCA [13, 20] is a system developed for preforming annotations by non-experts annotators. However, they experience UX problems where users struggle on mitigating the complexity of the system. The work carried out by Snow et al. [45] explored the use of Amazon Mechanical Turk to determine whether non-expert annotators can provide reliable annotations. They designed five different NLP tasks, among them NED and WSD, and for each of them they measure the quality of annotations by comparing them with the expert annotations. However, crowdsourcing is proved to be not an ideal solution to these type of problems [16].

3.2.3 Context Representation

In NLP, defining the context of the document or the surrounding context of a word, phrase or entity is generally seen as a high complex problem. Regarding NED and WSD, several research studies have used the *bag of words* model to measure the context similarity and consider this measure as an important feature to finalize the disambiguation decision [8]. However, according to Shen et al. [8], the bag of words model fails to capture various semantic relations existing between concepts. The bag of words model is nothing more than a vector representation of the context which consists of terms occurring in the window of text and their associated weights [8].

Other research studies go beyond extracting local contextual features to extracting context features from external sources such as Wikipedia pages. Cucerzan and Silviu [28] used the information present in the entities' Wikipedia page and other articles in which the entity is explicitly mentioned. Their strategy of representing the context of an entity is based on two category of references: first being the information present on the first paragraph of the target entity page and second being the starting paragraph of other entity pages which refer back to the target entity [28].

Unlike the others, the study conducted by Chan and Samuel [26] propose a semi-supervised approach for generating context templates to tackle the WSD problem. They have used a classification algorithm called Latent Dirilecht Alloction (LDA) which represents different topic features in a form of a vector space. All the feature vectors of the ambiguous word are recast into a network model. The disambiguation process is then done by calculating pairwise similarities between the context encoded in the templates (network) and the sentence of the ambiguous word, and taking the maximum value as the correct sense for the ambiguous word (please refer to [26] for a detailed explanation). A somewhat similar approach was also used by Navigli and Roberto [4]. They represent context using similar lexical features such as: tokenization, POS tagging, lemmanization, word chunking and parsing. Each target word is represented as vector of features, including the context features. The vector is used as a metric for the disambiguation step in the automatic algorithms.

Furthermore, linguistic features are proven to have a significant affect on the disambiguation of ambiguous entities. Zhou et al. [35] found that nouns are more informative than verbs by around 0.3%. They argue that nouns contain more contextual information than verbs because named entities are more salient (important).

Another important element to keep in mind when deciding on how to represent contextual information is the size of the context window. After conducting a user experiment, Bontcheva et al. [46] show that exposing only the sentence where the entity appears is not sufficient. It must be noted that the dataset they conducted the experiment with, was from tweets. Showing the whole tweet instead of the sentence resulted in better improved accuracy. It can be argued that when dealing with tweets, it makes sense to show the complete text as contextual information considering the short nature of tweets (maximum 130 characters). However, for long documents, showing the whole paragraph or one preceding and one following sentence (as suggested by Bontcheva et al. [46]) might degrade user experience and overwhelm the user with a lot of information to process.

3.3 Framework Architecture

The implemented NED Framework provides the fundamental tools and techniques on top of which a gamified system was implemented. Our ultimate goal is to demonstrate that games which are based on theoretical models and design principles can prove to be efficient in gathering qualitative annotations with minimal costs while still maintaining a large user base. Since the framework is composed of many different modules which carry various processing task and are independent from each other, an architectural model that adheres to these principles has been followed during the implementation. In other words, a micro-service architectural model has been utilized for the implementation of our framework. This section describes thoroughly all the different components composing the framework and the communication infrastructure used to exchange information between component.

3.3.1 The micro-service model

Implementing a complete monolithic architecture was seen as an unreliable and inconvenient solution for the nature of our problem. Besides the fact that many enterprises have started to shift from monolithic to micro-service architecture, one of the reason we decided to use a micro-service architectural style is the loose-coupling of components. Ceccarelli et al. [47] supports the idea that for an entity linking process a unique framework is shared where the recognition, disambiguation and linking processes are well separated and easy to isolate in order to study their performance. A microservice architecture is best fit for our case. According to Lewis and Fowler [48], a microservice architectural style is an approach for developing a complete application as a suite of small services, which are executed individually, each on its own process and communicate with each other using lightweight mechanisms, usually over HTTP. Implementing our framework in such a way allows us to follow a more generic and abstract approach to NED. It is generic in the sense that the different modules composing the framework can be easily changed to fit for other NLP tasks such as WSD, co-reference resolution or even changing the language of the task to something other than English. Illustrated in Figure 2, our framework is composed of 7 different microservices loosely coupled from each other with various responsibilities that build up the complete solution to the NED problem.

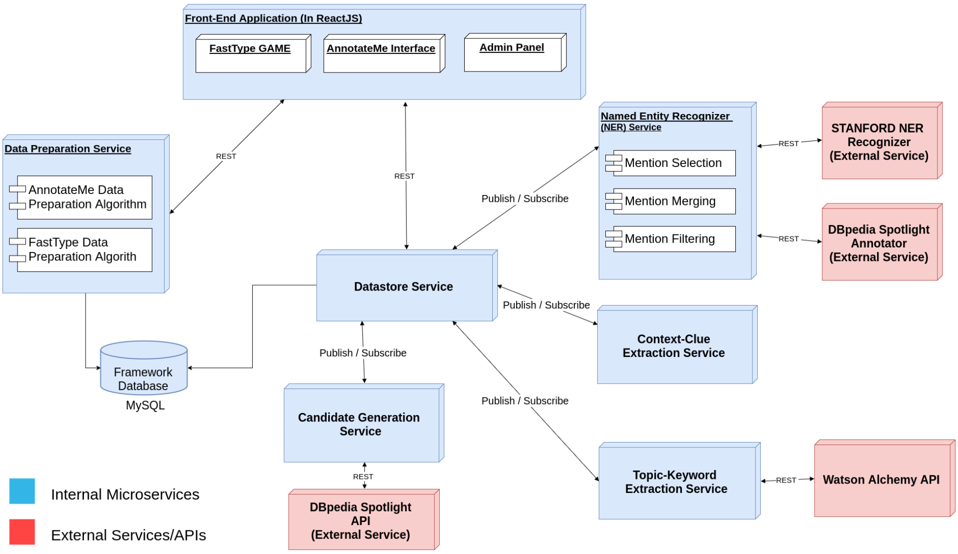

Front-End Application (In ReactJS)
FastType GAME
AnnotateMe Interface
Admin Panel
Data Preparation Service
AnnotateMe Data Preparation Algorithm
FastType Data Preparation Algorith
REST
REST
Publish / Subscribe
Named Entity Recognizer (NER) Service
Mention Selection
Mention Merging
Mention Filtering
REST
REST
STANFORD NER Recognizer (External Service)
DBpedia Spotlight Annotator (External Service)
Datastore Service
Publish / Subscribe
Context-Clue Extraction Service
Framework Database
MySQL
Publish / Subscribe
Candidate Generation Service
Publish / Subscribe
Topic-Keyword Extraction Service
REST
Watson Alchemy API
REST
DBpedia Spotlight API (External Service)
Internal Microservices
External Services/APIs

Figure 2: Overview of the NED micro-service architecture

Observing the figure above, two different categories can be identified. The internal microservices (boxes highlighted with blue) are services that have been completely implemented from scratch. On the other hand, the external microservices (boxes highlighted in red) are external open source services that have been used by our framework to perform actions which were out of scope of this study. Among the external services, *Stanford NER Recognizer* was the only external service that was not available as an online service offering API calls to perform the corresponding actions. A detailed explanation of the NER microservice is provided later in section 3.3.2.

In a microservice architecture, a key factor that differs this architecture from other design patterns is the lightweight communication/messaging infrastructure. In our framework, we have utilized two different communication protocols, namely, Synchronous Messaging (REST) and Asynchronous Messaging (AMQP). In cases where immediate response is required from one service to another, RESTful request and response API calls have been used to exchange information. We have tried to adhere to the fundamental rules of the RESTful messaging protocols where every functionality of the service is represented with a resource and operations carried out on top of these resources. As illustrated in Figure 2, the front-end service which contains the admin panel, AnnotateMe Interface and Fastype Game communicates synchronously with the data preparation service and data store service. Additionally, all external services are invoked using synchronous REST API. This is because our internal services are not able to perform any internal actions unless the response from the external service is acquired. A complete documentation of the REST API calls implemented in the data store and data preparation services are available in Appendix A.2.

Some of the services that compose the framework usually have a longer processing time compared to others because of the underlying complexity and the calculations that need to be performed. In these cases, it is more practical to use asynchronous messaging protocols without having to freeze the overall process as a consequence of one service which takes longer to complete. The publish and subscribe asynchronous message communication model was used for this purpose. More specifically, RabbitMQ[6] was used as the underlying lightweight messaging technology based on AMQP (Advanced Message Queue Protocol). To see the different publishing and subscription routes that build the asynchronous communication infrastructure of the framework, see Appendix A.1.

When designing a microservice architecture, among the many important design patterns that distinguish this architectural style from others is the deployment process. The deployment of microservices plays a critical role and when it comes to microservice architecture, the following key requirements have to be satisfied [49]:

- The ability to deploy/un-deploy independently each service
- It must be scalable at each microservice level (as some services may experience more traffic than others)
- The ability to build and deploy microservices quickly
- In case one microservice fails to execute, other microservices should not be affected by this failure

To comply with the above mentioned requirements, Docker[7] was considered as the best microservice deployment solution. Docker is a containerization tool that lets developers and system administrators deploy self-sufficient application containers in Linux environments [49]. Deploying a microservice into a docker container is as easy as writing a 5 line script. The steps involved to deploy an application into a docker container are as follows [49]:

- Packaging the microservice as a Docker container image (usually by writing a script)
- Deploying each service instance as a container
- Linking containers with each other so that they are able to communicate (this is done automatically by Docker)
- Scaling is done by deploying many instances of the same container

[6] RabbitMQ Documentation Page https://www.rabbitmq.com/documentation.html

[7] Docker Documentation Page https://docs.docker.com/

- Building, deploying and starting a microservice is relatively fast as Docker uses containers instead of virtual machines (which is much slower compared to containers)

In terms of technology-stack used to implement the microservice framework, the latest web technology tools have been utilized. In addition to the previously mentioned tools used for communication and deployment, the actual microservice applications framework has been implemented using NodeJS[8] as a back-end programming language, whereas for the front-end implementation we used ReactJS[9]. The combination of NodeJs and ReactJS has proven to be very efficient in terms of development speed, performance, agile development support and the incredible fast rendering capabilities which is very helpful during development and debugging of the application. As for the data storage, MySQL relational database has been used in our framework.

Figure 3 illustrates the complete workflow of the framework. We use this illustration as a reference when describing the different services composing the framework in the following sections.

[8] NodeJS `https://nodejs.org/en/docs/`

[9] ReactJS `https://facebook.github.io/react/`

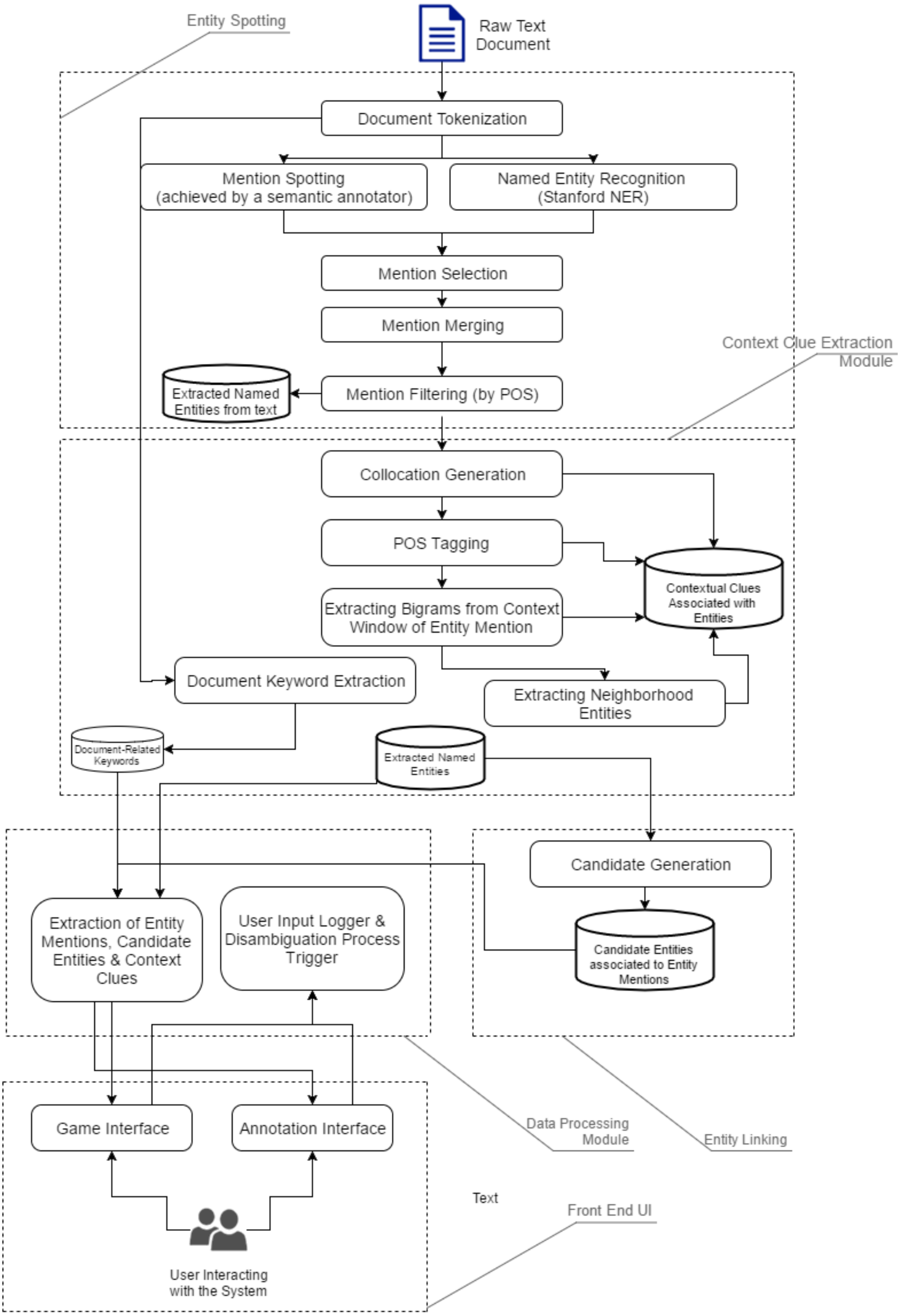

Entity Spotting
Raw Text Document
Document Tokenization
Mention Spotting (achieved by a semantic annotator)
Named Entity Recognition (Stanford NER)
Mention Selection
Mention Merging
Extracted Named Entities from text
Mention Filtering (by POS)
Context Clue Extraction Module
Collocation Generation
POS Tagging
Contextual Clues Associated with Entities
Extracting Bigrams from Context Window of Entity Mention
Document Keyword Extraction
Extracting Neighborhood Entities
Document-Related Keywords
Extracted Named Entities
Candidate Generation
Extraction of Entity Mentions, Candidate Entities & Context Clues
User Input Logger & Disambiguation Process Trigger
Candidate Entities associated to Entity Mentions
Game Interface
Annotation Interface
Data Processing Module
Entity Linking
Text
Front End UI
User Interacting with the System

Figure 3: Overview of the complete workflow of the NED Framework

3.3.2 Named Entity Recognition (NER) Service

The disambiguation process is initiated by uploading a raw text document through the admin panel which issues an API call to the data store service. The data store service persists the content of the document into the database and proceeds by publishing a *named entity recognition* message which in turn is subscribed by the NER microservice. The published message contains all the textual content of the uploaded document. As can be seen in Figure 3, the NER microservice starts by tokenizing the textual content of the document. A NodeJS implementation of a tokenizer[10] has been used for this purpose. The tokenization process converts the text content into an array of words, sentences, characters or any other desired way. We use the tokenizer to split up sentences and individual words.

The whole named entity recognition procedure has been inspired by [43], as they achieve very high levels of accuracy and significantly outperform state-of-the-art named entity recognizer. However, the implementation of the NER service by Chabchoub et al. [43] was carried out as part of the OKE Challenge 2016 and we were not able to find any open source code that could be integrated into our framework. Based on the information the authors provide in their publication, an attempt was made to reproduce the algorithms used in the recognition module in order to achieve similar performance levels as reported in their publication [43].

Stanford NER [38] has been used as a named entity recognizer in combination with Dbpedia Spotlight [10] as a semantic annotator. Both (as external services) recognize the entities in the textual content and return a list of all the recognized entities. When comparing the results of each service, entity overlapping was observed. To be able to solve this, a selection algorithm is implemented in order to select the best entity out of both lists. The logic of the selection algorithm is keeping the longest mention and dismissing the short one. Lets take an illustrative example from the sentence below:

> *"The State University of New York at Cortland celebrated its 149th anniversary this year."*

In this example, Spotlight annotates *State University* and *New York* separately, whereas Stanford NER recognizes it as a single entity, namely, *State University of New York*. The mention selection algorithm makes sure that the former is discarded and the later is kept.

However, even after the mention selection algorithm, it is not guaranteed that the correct entities have been identified. From the example sentence, in fact, the correct entity mention is *State University of New York at Cortland*. In order to achieve this, the mention merging algorithm is performed. As explained by [43],

[10] See Node Tokenizer https://www.npmjs.com/package/tokenize-text

given two named entities in close proximity, the algorithm will try to expand it to cover the next entity mention. The constraints checked by the algorithm to permit the expansion (avoiding pitfalls such as merging two legitimate different entities) have been described in detail by Chabchoub et al. [43] and used as a guide for implementing the logic of the mention merging algorithm.

The last step, before the entities are persisted into the database, consists of applying the mention filtering algorithm to the list of identified entities. The mention filtering algorithm uses a standard Part-of-Speech tagger for getting linguistic information for each entity. Accordingly, all entities that contain verbs are removed from the list, thus, filtering out incorrectly recognized entities [43].

The final result of the NER service contains a list of *selected*, *merged* and *filtered* entities. The NER service finalizes its process by publishing a *named entity persist* message which is subscribed by the data store service that does the actual persisting of the entities into the MySQL database.

3.3.3 Context Clue Extraction Services

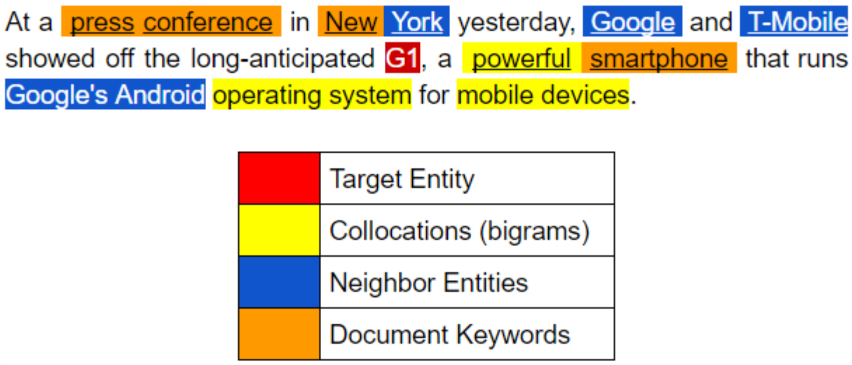
During the background chapter we emphasized the importance of context and how much it can affect the disambiguation accuracy for both, automatic and manual annotation systems. Referring back to the work-flow presented in the diagram in Figure 3, contextual clues are extracted by following a three step process.

The process starts by removing all stop words in the sentences where the entity mentions are part of. After the stop words have been removed from the sentences, the *collocation generation algorithm* is performed on those sentences. Collocations, according to Colson et al. [50] represent the occurrence of two or more words within a short proximity between each other in text. They also argues that collocations are more likely to occur as fixed expressions such as compound nouns, proper nouns, idioms, noun-adjective combination, adjective-noun combination, verb-noun, well known song or film titles. The collocation generation algorithm follows these principles when deciding to keep a collocation from the sentence or not. Since we are dealing with analysis of grammatical structures in the sentence, part-of-speech tagging is a crucial step to be performed at this stage.

After having extracted all potential collocations from the sentence provided as input data to the service, the next step is to associate these collocations as contextual clues to each and every entity. However, not every identified collocation within the sentence can be a useful context clue for the entity that is also part of that sentence. Previous studies have used a context window size of 4, which means that they take four preceding and four subsequent words from the sentence where the target entity occurs [31]. The number 4 has been chosen because the accuracy of sense resolution does not improve when more than 4 words around the target word are considered. However, recent studies argue that the context windows size around the ambiguous target entity is dependent on the nature of the word itself [31]. Our solution to this problem is increasing the context window size depending on how ambiguous the target word is. The ambiguity of a target entity is defined by the number of potential candidates extracted from the KB. The more candidates are generated for the target entity, the higher the level of ambiguity. We believe that this represents an accurate metric on deciding the context size of the target entity.

Since collocations can be a combination of more than two words, we have decided to keep only collocations that are composed of two words (i.e. bigrams). This decision is influenced on the findings acquired by Mihalcea and Rada [36]. They observed that in terms of words in context, bigrams seem to be more effective than simple keywords and other combinations larger than two. In addition to the collocations which are considered as clues to help summarize the context in which the target word occurs, neighbour entities also fall into this category. Since the algorithm has information on the exact position of each entity in the sentence (by maintaining a starting and ending index), we are able to decide which (other) entities are in close proximity to the target entity. Usually all identified entities that are part of the same sentence as the target entity are considered as neighbor entities. In some cases, when the sentence is short, the algorithm looks for neighbor entities in the preceding and subsequent sentences. The algorithm maintains a constraint on which it bases its logic whether to consider keeping or discarding a neighbor entity for the target entity. The constraint is basically a calculated word distance between the target entity and all other potential neighbor entities.

After the local contextual clues have been extracted by our internal microservice, the last step is the extraction of document keywords. Document keywords are used to represent the theme or topic of the document, and might help in the disambiguation process in addition to local contextual clues. In this stage, Alchemy API is queried which extracts the most important keywords from a document. These keywords are associated as context clues to all identified entities in the same document whereas local context clues are distinct from one entity to another. Figure 4 illustrates an example of a paragraph and the corresponding results after running the context clue extraction service. In this example, the target entity for which the context clues are to be associated is highlighted in red. The legend illustrated as a table in the example figure explains the meaning of the different colors. In cases where the words are underlined and highlighted with different colors, it means that the word combination has multiple purposes. For example *New York* from the illustrative paragraph in Figure 4 is considered a neighbor entity to G1 (the target entity) in addition to being a keyword that contributes to understanding the



At a press conference in New York yesterday, Google and T-Mobile showed off the long-anticipated G1, a powerful smartphone that runs Google's Android operating system for mobile devices.

🟥	Target Entity
🟨	Collocations (bigrams)
🟦	Neighbor Entities
🟧	Document Keywords

Figure 4: An example of a paragraph and the different contextual clues extracted from the service

meaning of the complete paragraph.

3.3.4 Candidate Generation Service

Dbpedia Spotlight is the semantic annotator which is queried by the *Candidate Generation Service* to extract Dbpedia candidates for a specific entity mention. No special logic has been applied in this service and therefore, the results retrieved after querying Dbpedia Spotlight are persisted into our system without additional modifications to the original data.

Dbpedia Spotlight [10] performs the disambiguation process by pre-ranking entity candidates for each surface form spotted in the text. A combination of a prior score and a contextual score is calculated in order to determine which candidate entity is the most relevant. The prior score represents an estimation of how often the surface form is used as an anchor in a Wikipedia hyperlink that points to the entity page [43]. Whereas the context score makes use of the context of the phrase (usually a window of words around the phrase) and the context of each candidate entity (calculated internally by Spotlight). When querying highly ambiguous entities such as *Paris* which can have over ten target candidates, only the best 8 are fetched to be evaluated. According to a user study conducted by Bontcheva et al. [46], participants gave feedback that having more than 8 options to choose from is associated with high cognitive load and results in immediately exhausting users. In addition to the maximum number of 8 candidate entities, we provide a last option called *none of them* in case the Spotlight was not able to fetch the correct candidate in the list.

During the experimental user study we observed that Spotlight was not able to provide the correct candidate entity for many entity mentions. Therefore we experimented with providing different context window sizes to the annotator to see if the performance changes. First we queried Spotlight by providing only the entity itself without any other contextual information. Second, the contextual clues extracted by our microservice were put in a sentence together with the target entity with clues located prior and after the target entity (based on where the clues were located on the original sentence). Finally, the original sentence where the target entity is part of, was used as context and sent as query parameter to Spotlight. We report in the results section of this chapter that the differences in performance between the three groups is relatively small and does not assure statistical significance.

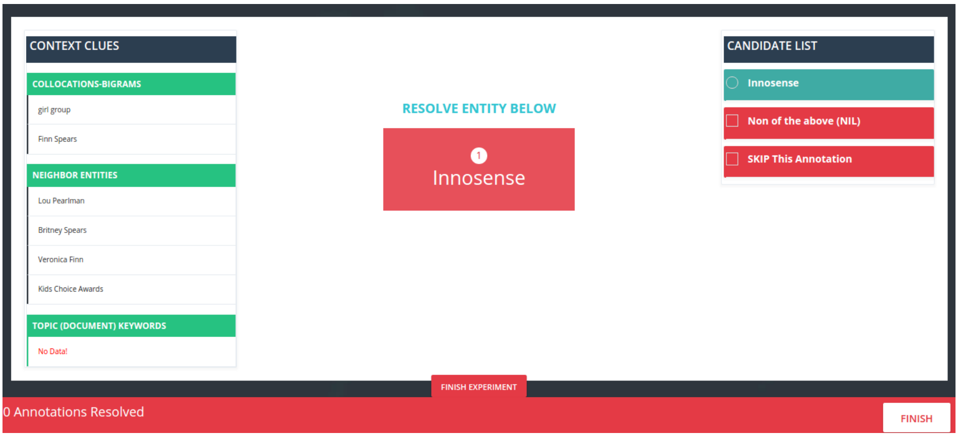
3.3.5 Data store & data preparation Service

To assure the control and consistency of data being processed by the different microservices composing the framework, the *Data-Store Service* was designed to be the only entry point to which the data could be manipulated. The service provides endpoints for accessing and manipulating the information residing on the database as a means of API calls and asynchronous message inquiries. It also serves as an information provider to the admin panel in the front-end application and also subscribes to different message routes such as: persisting named entities recognized from NER microservice, persisting context clues and associating the generated candidate list to all registered entity mentions in the database. Besides subscribing to these message routes, the *Data-Store Service* is the service endpoint that manages the complete asynchronous messaging infrastructure.

On the other hand, the *Data-Preparation Service,* as the name implies, prepares the data for the AnnotateMe Interface as well as for the Fastype Game. Similar to the Data-Store Service, it has direct access to the database information with only one specific permission: reading (i.e. querying and retrieving information from the database). Therefore, for the sake of centralization and control of data, the Data-Preparation Service is considered as a read-only service with regards to the database access.

A unique feature implemented in the data preparation service is, as we like to call it, the *disambiguation trigger*. This feature is responsible for resolving a specific entity mention (deciding which candidate represents the correct link for the target entity) when enough annotation data from the human annotators are accumulated. Since the most usual use-case scenario includes non-expert human annotators performing the validation process by either using the AnnotateMe Interface or the Fastype game, assuring quality of annotations is reached through redundancy. The level of redundancy maintained in the data preparation service is based on constraints proposed by Snow et al. [45]. They conducted an experiment where they evaluated the quality of non-expert annotators in comparison with expert annotators. Their results indicate that on average it requires 4 independent non-expert annotations to achieve the equivalent ITA of a single expert annotator. Therefore, the disambiguation trigger is triggered when 4 independent annotations (having the same candidate as the selected option) have been accumulated for a specific entity mention. After an entity has been resolved, it will no longer show up on the interface for validation. The resolving step is done by the *Data-Preparation Service* which initiates a REST API call to the *Data-Store Service* in order to update the information on the database.

A final element that is taken care by the *Data-Preparation Service* is the ordering of candidates presented to human annotators. In a study conducted by Duarte et al. [51], they argue that in search engines, web users expect the best answer to be in the first or second position. This type of expectation represents potential bias on the results assessing the behaviour of annotators. After performing some user studies, they conclude that search result selection behaviour is influenced by ranking, with users showing tendency to select higher ranks without exploring other alternatives. To avoid having this situation in our experiment, we encourage users to explore all the candidates presented to them before making a decision. The encouragement is done by providing short descriptions for each candidate on the interface. Additionally, we avoid the chance of making users form assumptions about potential candidates being ranked higher in a list by completely randomizing the process of candidate positioning. Random positioning of alternatives instead of ranking has proven to be much more effective in encouraging users to explore all available alternatives instead of making blind decisions [51].



CONTEXT CLUES

COLLOCATIONS-BIGRAMS

girl group

Finn Spears

NEIGHBOR ENTITIES

Lou Pearlman

Britney Spears

Veronica Finn

Kids Choice Awards

TOPIC (DOCUMENT) KEYWORDS

No Data!

RESOLVE ENTITY BELOW

1

Innosense

CANDIDATE LIST

○ Innosense

□ Non of the above (NIL)

□ SKIP This Annotation

FINISH EXPERIMENT

0 Annotations Resolved

FINISH

Figure 5: The AnnotateMe Interface used to conduct the first experiment for validating the links generated by the NED Framework

3.3.6 AnnotateMe Interface

During the implementation of the AnnotateMe Interface, we tried to come up with a simple UI and UX design of the interface in order to maintain low levels of complexity and avoiding confusion. Some of the design patterns identified by Hinze et al. [13] necessary for non-expert users to annotate the presented content have been explored while implementing the annotation interface. These design patterns include:

- Intuitive User Interface - an interface that is easy to grasp with actions that require minimal effort to discover and perform
- Simple Vocabularies - the current architecture of the framework provides annotations of entities within the categories of Organization, Location and People which are genuinely simple in nature.
- Focus on user task - The interface does not have any disruptive features or elements that would shift the focus of the annotator from its main task, that is, resolving the presented named entity.

According to Bontcheva et al. [46] the single most influential part of any linguistic annotation exercise is the annotators ability to understand and conduct the annotation task. In order to achieve this, the guidelines and tools provided by the annotation interface play a major role in controlling such behavior. Presenting the user with simple, short guidelines that include examples and specific instructions on how to perform certain actions is very helpful. Additionally, having a clean interface is of utter most importance which contributes to an intuitive interaction during the annotation process. [46]

Figure 5 represents the design of the AnnotateMe Interface used for conducting annotations with non-expert users. The target entity mention to be resolved is presented in the middle of the page and attracts the users focus by reflecting its importance in the overall task. The contextual clues are presented to the right hand side of the interface and are grouped based on their origin of extraction. The right hand side of the interface is reserved for the candidate list. Each candidate can be expanded by clicking on its name. The expanded candidate provides the user with additional information that describes the meaning of the candidate (usually a short description). The two last options in the candidate section highlighted with red provide some degree of freedom to the user in case they are unfamiliar with the entity or when they think that the correct candidate is not provided in the list. These two options being discussed are: *Non of the above(NIL)* and *Skip this annotation*. The interface presented in Figure 5 has been used to conduct the first experiment which will be described in detail in the next upcoming section.

3.4 Methodology

The implementation of the Named Entity Disambiguation Framework with all its underlying microservices represent the tool which is used to study and answer the first two research questions. The majority of the implementation patterns and algorithms are based on theoretical background and reviewed literature on the respective field. In absence of open source code, the NER service has been implemented as a reproduction of the work conducted by Chabchoub et al. [43]. However, without actual contribution of human annotators by participating in a user study, the answers to the research questions would not have been resolved. The first experimental user study has been conducted in order to get detailed insights on how well the information processed and presented to human annotators would contribute to generating annotation data. The reason that two user studies were conducted during this research work was to measure the engagement and playfulness of the game when compared to a standard, non-gamified interface for named entity disambiguation. Data from both experiments have been used to analyze and compare the different characteristics in order to draw on conclusions that we initially set to achieve. Regardless, this methodology section represents all the necessary information needed to successfully conduct the first experiment and describe the techniques and methodologies used to analyze the results.

In order to conduct the annotation experiment by using the designed AnnotateMe Interface, an annotation data gathering process was conducted beforehand. In absence of linguistic and expert annotators to construct a gold standard which would have been used for assessing the quality of annotations, we decided to use already existing datasets, namely, Spotlight and KORE50 datasets [3]. The first dataset contains documents from news articles whereas the second one contains short articles extracted from various domains. KORE50 is known by the NLP experts as a dataset characterized with a highly ambiguous nature. Table 1 summarizes all the different entity types that are recognized in each dataset used in our evaluation experiment. As observed from the table, both datasets are composed mostly of entities in the three main categories: Organization, Location and People. The table was originally reported by Steinmetz et al. [3], and during the time of their writing, some of the entity mentions recognized in the dataset did not have a KB representative. However, it has been mentioned that the information in LOD is continuously increased by researchers contributing with new datasets and extending existing ones. Therefore, the ratio of mentions and entities from both datasets is likely to be more equalized now[11]. However, the analysis and results presented in the next section are up-to-date and take into consideration the latest versions of KROE50 and Spotlight datasets.

[11] The number of entity mentions recognized in the text having a corresponding KB representative is increased.

Table 1: Distribution of entity types for Spotlight and KORE50 [3]

Class	Spotlight		KORE50	
	Entities	Mentions	Entities	Mentions
Total	249	331	130	144
Agent	1.40%	2.70%	66.90%	70.80%
-Organization	<1%	<1%	19.40%	5.30%
–Company	<1%	<1%	9.20%	9.70%
–Sports Team	-	-	7.70%	6.90%
—Socer Club	-	-	7.70%	7.90%
-Person	2.00%	2.40%	48.50%	51.40%
–Artist	-	-	17.70%	18.80%
—MusicalArtist	-	-	17.70%	18.80%
–Athlete	-	-	6.90%	8.30%
—SoccerPlayer	-	-	5.40%	6.30%
–OfficeHolder	<1%	<1%	4.60%	4.20%
Disease	1.60%	1.20%	-	-
EthnicGroup	1.20%	1.80%	-	-
Event	1.20%	<1%	-	-
Place	10.40%	10%	10.80%	10.40%
-Architectural Structure	2.00%	1.50%	3.10%	2.80%
–Infrastructure	1.60%	1.20%	<1%	<1%
-PopulatedPlace	7.20%	7.60%	5.40%	5.50%
–Country	3.60%	3.30%	-	-
–Region	<1%	<1%	-	-
–Settlement	2.40%	3.30%	3.80%	3.50%
—City	1.60%	2.10%	2.30%	2.10%
Work	<1%	<1%	6.20%	6.30%
-MusicalWork	<1%	<1%	3.10%	3.50%
–Album	<1%	<1%	3.10%	3.50%

The participants who took part in the annotation experiment were invited through emails and social media. On the invitation message sent to all participants, a short and abstract description about the idea of the experiment was provided. It is important to note that payment incentives or any other type of incentive which would degrade the voluntary participation of the user were avoided. Thus, we assure a complete voluntary participation without any beneficial intentions. As a result, 30 participants showed up and successfully completed the experiment.

The experiment was conducted in closed group rooms at the university campus in order to make sure that the participant was not being distracted while performing the experiment. Before starting to perform the actual annotations, the participants were presented with a consent form which explained the nature, purpose and intentions of the experiment. Participants were also fully aware that the participation was anonymous and voluntary and withdrawal from participation was possible at any time. The consent form for the first experiment is provided in Appendix A.

On average, the duration of a single experiment session was about 25 minutes long. We also asked the participants to fill in a pre-questionnaire to gather demographic information about them. Additionally, questions that assessed the level of expertise of each participant in the field of semantic web and natural language processing were posed in this questionnaire. In order to avoid potential biases on the results, we made sure that participants had moderate to native skills in English. We report on the demographic data of participants in the next section. Questions provided in the pre-questionnaire can be found in Appendix A.

The non-expert nature of participants and the time constraints on the duration of an experiment session was the main reason for recording an instructional video for demonstrating the usage of the interface to the participants. This video was essentially a replacement for an introductionary stage that is usually implemented for such experiments. During the 4 minute video, participants were instructed on how the interface is used and the purpose of each elements towards the general idea of annotating. After having read the consent form, filled in the pre-questionnaire and watched the demo video, the participants proceeded in doing the annotations for 15 minutes. However, the experimenter did not provide upfront information on how many annotations had to be performed. After the estimated 15 minutes elapsed, the participants were instructed to either finish the experiment or continue do more annotations. This is a technique we used for measuring the level of enjoyment that participants had towards the interface. After the participant free-willingly[12] decided to finish the experiment, he or she was asked to fill in a post-questionnaire. The post-questionnaire contained questions with the purpose of assessing the quality, usability and engagement level of the interface. We report on the results of the post-questionnaire in the next section. The questions provided in the post-questionnaire can be found in Appendix A.

For assessing the performance of the framework in terms of entity recognition, user annotation quality and agreement level, we used metrics such as precision, recall, f-score and one way ANOVA analysis of variance. These assessment methodologies are commonly used in entity linking systems [5]. Furthermore, to be able to assess the performance of our approach with state-of-the-art frameworks we have used a generic benchmark called GERBIL originally developed by Usbeck et

[12] A free-will annotation is an annotation performed by a participant after being aware that the experiment was completed. This is a metric to measure the attractiveness of and engagement with the interface as perceived by the participant.

al. [52]. GERBIL is a comparison tool for easily discovering the strengths and weaknesses of your implementation with respect to the state of the art in named entity disambiguation. The tool is open source and is an extensible framework that currently supports 9 different annotations on 11 different datasets within 6 different data types (recognition, disambiguation, linking etc). We report on the results of our framework with respect to the state-of-the-art annotators offered by GERBIL for Spotlight and KORE50 datasets.

Comparing the performance results of our framework with state-of-art automatic supervised approaches for NED represents another important design decision that shifted the direction of this study in the way it is. One might argue that comparing annotation performance of human annotators with performance of supervised algorithms does not represent a fair comparison. However, the initial idea of this research originates from the fact that the lack in availability of expert annotators to generate large-scale annotation data motivated this study to take this specific approach. Supervised approaches require an enormous amount of training data in order to improve their accuracy and performance for annotation. Since expert-annotators are not able to generate such data, the knowledge of the sheer Internet user has to be leveraged with appropriate techniques for this purpose. Therefore, acquiring performance results for annotations from ordinary non-expert users which are significantly better compared to the performance of supervised approaches, we are able to claim that the generated data can be used for improving the performance of the later. Consequently, we argue that our selected comparison and assessment methodology is appropriate for investigating and finding solutions to the problem addressed in this research.

Concerning limitations and potential sources of bias, the methodology we used for our research study can be considered partially immune. For the sake of reproducability, we need to note that the participants who were invited through emails and social media were within the scope of the university campus. Consequently, one might argue that they might have been biased to participate in the experiment. Since this was seen as the only way of recruiting the participants in the pressure of time and space, we account this as a potential source of bias and address as a limitation of the recruitment methodology. However, we assure consistency with regard to participant recruitment since the same approach was used to recruit participants for the second experiment as well. Therefore, we are able to claim that the improvements are acceptable since we remain consistent in the recruitment methodology.

Hypothesis

The aim of the first experiment is to find out whether the implemented framework supports qualitative annotations by non-expert annotators. Therefore, we hypothesize that:

- H1.1: By implementing the complete entity disambiguation framework, non-expert users will be able to perform high quality annotations with an observed improvement compared to supervised automatic approaches!
- H1.2: Short contextual clues such as bigrams, neighbor entities and topic keywords are preferred towards complete sentences or paragraphs and as such provide sufficient information to make correct disambiguation!

3.5 Results

The reason for performing the first experiment with the AnnotateMe Interface was to find out whether our approach of formulating and presenting the annotation data to a non-expert annotator for validation would result in creating high quality annotations. During this experiment we have also tested the performance of our NER service to see how well the recognition task is performed. In addition to that, the performance of Dbpedia Spotlight as the utilized automatic annotator was also tested. Below we report on the different aspects of the experiment.

3.5.1 Participants

The non-expert annotators who took part in the first experiment can be characterized as moderate users of text processing tools who have (on average) moderate experience with tagging of textual documents. Please note that tagging was explained to the users as the process of assigning a label or a textual description to a picture, video or categorizing a document. In some cases, the easiest way of explaining the tagging process to a participant was taking the concept of a *hashtag* as an analogy to tagging. However, a *hashtag* provides a higher degree of freedom since the tagging process is open while in our case the user is presented with options to choose from. Regardless, they both share the same conceptual idea. These questions were used to obtain a rough idea about the informal level of expertise of each participant in text processing respectively.

The average annotator was a 25 years old male student studying in the field of Computer Science with a good familiarity of text processing tools, moderate experience with tagging, acceptable level of familiarity with semantic web concepts who considered himself as above average when asked for his English language skills. Figure 6 presents the results of the questions asked to the participants during the pre-questionnaire.

3.5.2 NER Performance

It has been mentioned earlier that the algorithms implemented in our NER microservice are not novel solutions as opposed to the state-of-art. We attempted to reproduce the same algorithms used by [43] since at the time of writing, their approach performed best compared to any other state-of-art NER algorithms and tools. However, in absence of open source code, we tried to implement the service as similar as possible based on the information the authors provided in their white paper submitted at the OKE Challenge 2016 [43]. As we can see from Table 2, our framework performs significantly better than all the other automatic annotators for the KORE50 dataset. For the Spotlight dataset however, we observe a slightly better improvement. Please note that we used GERBIL Benchmark [52] to get the values for the other annotators whereas for our framework we calculated the same

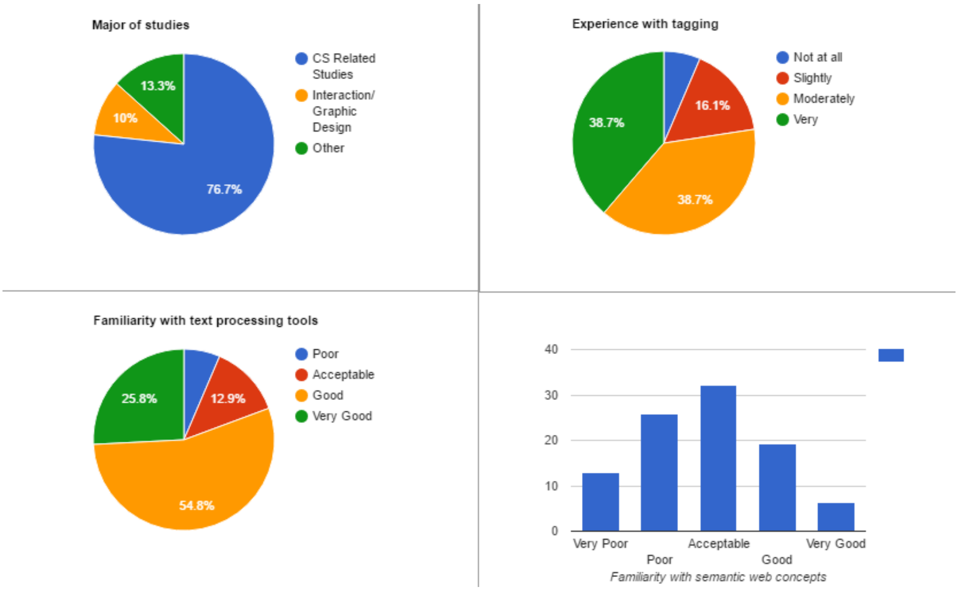


Figure 6: Participant Statistics for Experiment 1

measures as explained by the benchmark tool.

Table 2: NER Performance of automatic annotators in two datasets compared to our framework

-/Dataset	**Spotlight Dataset**			**KORE50 Dataset**		
	Precision	**Recall**	**F-Score**	**Precision**	**Recall**	**F-Score**
Babelfy	0.22	0.18	0.19	0.55	0.55	0.53
Dbpedia	0.61	0.34	0.43	0.73	0.23	0.27
AIDA	/	/	/	0.64	0.53	0.57
Dexter	0.71	0.27	0.38	0.27	0.15	0.18
AnnotateMe	**0.69**	**0.42**	**0.49**	**0.87**	**0.8**	**0.82**

Our intentions were to see how good our framework performs when dealing with entities that have a very high ambiguous nature (KORE50 dataset) but also with entities that have moderate levels of ambiguity (Spotlight dataset). However, the number of entities present in each dataset was relatively big and resolving all of those entities using the framework would require running the experiment for longer periods. Therefore we used only a subset of documents from the Spotlight dataset wheres for the KORE50 dataset we used all the available documents. For reproducability reasons, the following documents from the Spotlight dataset were considered: Arts1, Business1, Fashion1, Medicine1, Music1, Privacy1, Science1, Sports1, Travel1 and Travel2. The complete dataset is available online at the GERBIL website[13]. The total number of recognized entities by our NER microservice was 208 with 77 being entities recognized from the Spotlight dataset and 131 entities recognized from the KORE50 dataset.

3.5.3 Annotation Quality and Performance

Among the 208 total recognized entities by our NER microservice, only 82 of them were resolved during the first experiment. It should be emphasized again that for one entity to be resolved we required 4 unique judgments from different participants to agree on one specific candidate for it to be resolved. We measured the performance of the human annotators, Dbpedia Spotlight and other automatic annotators on those 82 resolved entities. Furthermore, we evaluated whether the candidate associated with the entity was the correct one based on the gold standard. Additionally, since we wanted to see how Dbpedia Spotlight (as our utilized automatic annotator) disambiguate entities based on the amount of contextual information surrounding the target entity, we report on three cases. The cases include: providing only the entity itself without any context information, providing the complete sentence and providing the contextual clues extracted from our framework.

Table 3 presents the ratio of correctly and incorrectly linked entities by Dbpedia Spotlight. In general, the performance increased after providing more contextual information to the automatic annotator, which is an expected outcome. However, against our expectations, in the case of Spotlight dataset, the ratio of correctly linked entities decreased after providing additional context information to the annotator. Despite the fact that the difference between the groups is not statistically significant for $p < 0.05$, it leads us to the assumption that Dbpedia Spotlight requires more content than just short contextual clues in order to make use of their internal techniques for disambiguation. Finally, we compare our performance results with other automatic annotators aside from Spotlight and observe a significant improvement on both datasets. Table 4 reports on the F-score of each annotator compared to our framework.

Additionally, we run one-way ANOVA analysis of variance between two groups, namely, annotation results performed by non-expert users using the AnnotateMe Interface and results obtained by Dbpedia Spotlight automatic annotator. Statistically significant differences for $p = 0.05$ are obtained. As a result we claim that the non-expert annotators performed significantly better compared to other semantic annotators with an observed accuracy of 0.92 for the f-measure. Based on these

[13] http://aksw.org/Projects/GERBIL.html

Table 3: Annotation results of AnnotateMe Interface and Dbpedia Spotlight annotator

Experiment 1 (Entity Name Only)			
	All Datasets	**Spotlight Only**	**KORE50 Only**
Correct (%)	44.57	72.22	32.23
Incorrect (%)	55.43	27.78	67.77
Experiment 1 (With Sentances)			
	All Datasets	**Spotlight Only**	**KORE50 Only**
Correct (%)	51.38	70.37	43.31
Incorrect (%)	48.62	29.63	56.69

Experiment 1 (With Context Clues & Neighbor Entities)			
	All Datasets	**Spotlight Only**	**KORE50 Only**
Correct (%)	54.24	66.67	48.78
Incorrect (%)	45.76	33.33	51.22
AnnotateMe			
Correct (%)		**Incorrect (%)**	
88.46		**11.53**	

results, the first hypothesis (H1.1) is strongly supported.

Table 4: Comparison of AnnotateMe with other state-of-art semantic annotators

Precision, Recall, F-Scores (Spotlight & KORE50 Dataset)			
Annotator	**Precision**	**Recall**	**F-Score**
AnnotatMe	**0.95**	**0.88**	**0.92**
Babelfy	0.56	0.46	0.51
Dbpedia Spotlight	0.53	0.39	0.44
Dexter	0.27	0.17	0.2
WAT	0.55	0.41	0.46

In order to answer our second research question regarding context, we asked participants in the post questionnaire whether the contextual clues were helpful during the annotation process. They were also asked whether they would prefer to have the complete sentence as a representative of the surrounding context of the entity or stick with the short context clues. The results from the post questionnaire analysis show that 66.67% of the participants preferred the short context clues as opposed to the 31.25% of the other group who said that they would prefer sentences instead. To see whether contextual clues really helped on the disambiguation process for those participants who approved them as helpful clues, we calculated the level of agreement of each participant. The level of agreement is a simple calculation that takes the number of correct entities resolved and divides it by the total number of annotations performed. Results show that from the group who agreed having short contextual clues rather than sentences, on average they agreed with other annotators 54.47% of the time. On the other hand, those who preferred sentences agreed on average 48.44% of the time. We observe a slight improvement on the agreement level in this case. Figure 7 presents the overall agreement levels of all participants who took part in the first experiment. The average agreement level for all participants is 51%. Please note that this does not represent a 50-50 chances of agreeing with another participant. In most cases, when resolving an entity, each participant was presented with 8 options in addition to the 9th option being *none of the above*. Therefore, a 51% reported agreement level can be considered a good performance. However, the difference between the agreement level reported for the group who preferred context clues compared to those who preferred sentences is not statistically significant for $p = 0.05$. As a result our second hypothesis (H2.1) is rejected.

The setup of the first experiment was designed in a way that each individual

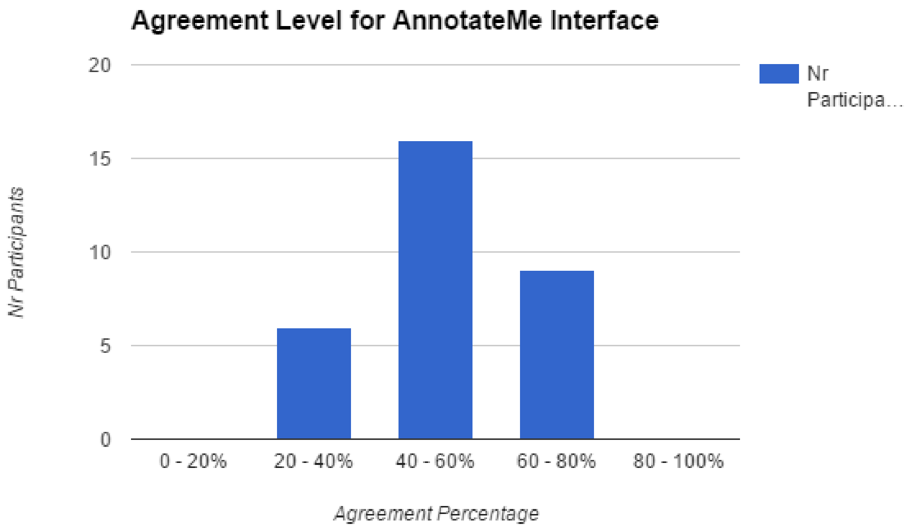


Figure 7: Participant's Agreement level for the first experiment

participant performed annotations for 15 minuets constantly. They were not told upfront that they will be performing such tasks continuously for 15 minutes. They were instructed to do the task until the experimenter told them to stop. After the estimated time elapsed, participants were asked whether they would like to continue perform annotations or stop and conclude the experiment. Participants were not encouraged to do any annotations after the 15 minute time trial had passed. Therefore, the rest of annotations performed were completely voluntary and considered as free-will annotations. On average, each participant performed 22 annotation rounds during those 15 minutes. In terms of free-will annotations, one participant performed 17 annotation rounds on average. These results are used later to compare with the gamified system in order to assess the perceived engagement as a measure of time spent performing *free-will annotations*.

On the post questionnaire we asked participants to evaluate their experience with the task, the usability of the interface and perceived engagement. Figure 8 reports the results on the different questions asked on the post questionnaire. It can be concluded that the participants perceived the task to be interesting and somewhat engaging with a possibility that the participants would continue perform such task in the future. Regarding the usefulness of contextual clues, the graph indicates that users perceived the clues on average as useful when disambiguating a named entity. In addition to that, we report also on the perceived frustration

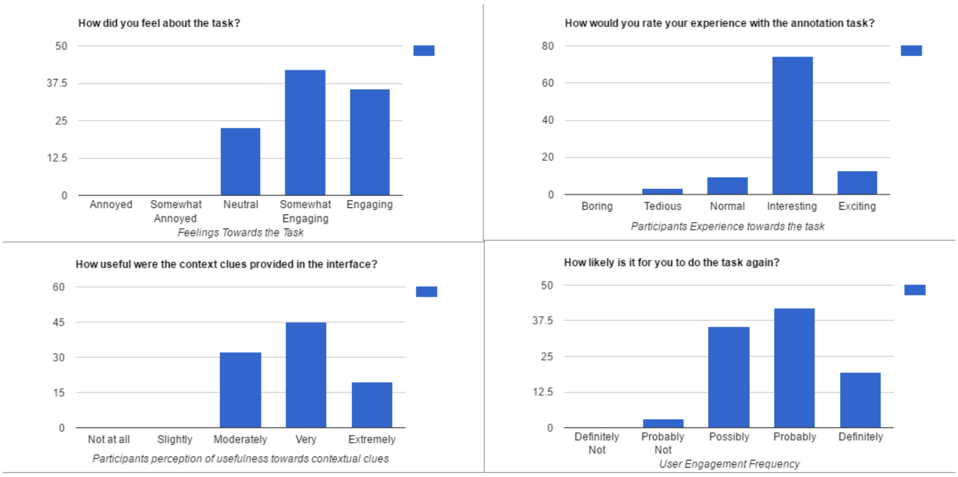


Figure 8: Experiment 1 Post Questionnaire Results

while performing the task. Figure 9 indicates that the participants were seldom frustrated during the annotation process with some of the participants reporting to having been frustrated about half of the time. This can be an outcome of the participants being not familiar with the entities since they did not have the freedom of choosing a specific category or genre during the first experiment.

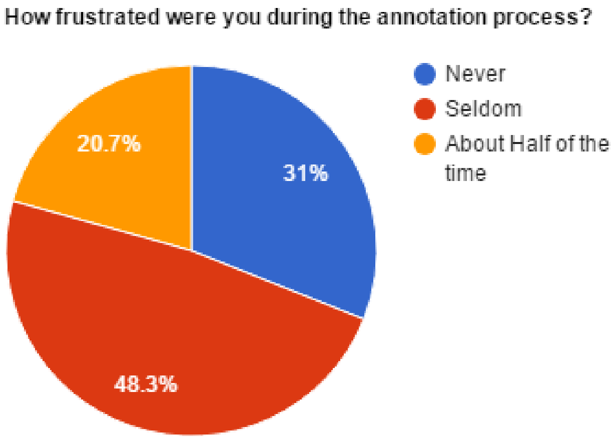


Figure 9: Reported level of frustration experienced during the annotation process

4 Fastype - Gamification of NED

In the previous chapter we presented the named entity disambiguation framework which has been implemented as the initial part of the work conducted for this research study. Results from previous experiment indicate that non-expert users are able to perform qualitative annotations with the help of short contextual clues and an intuitive user interface. However, results from using the plain interface also indicated seldom levels of frustration during the experiment in addition to a slight percentage feeling indifferent/neutral towards being engaged or positively motivated in performing the task. For the long-term run, we assume that the plain interface lacks elements of engagement and does not have any associated intrinsic motivation that will bring users back to perform annotations. We hypothesize that by applying appropriate game design to NED, participants will feel engaged and intrinsically motivated to perform annotations. In this chapter we provide an outline of related work in gamification and the idea of gamifying complex systems. We primary focus on gamification of systems in the field of semantic web and natural language processing (NLP). Background information on specific game design and theoretical models on which we base or work are also elaborated. Additionally, we explain the methodology used for implementing the game, preparing the data for the user study in addition to explaining the metrics and assessment techniques used to analyze the data and report on the results.

4.1 Background

Gamification is not about taking an existing system and decorating it with points, levels and leaderboards. This methodology of gamification is referred to as "Pointsification" where the game design exclusively relies on points, badges and leaderboards [21]. Zichermann et al. [53] argues that the technique of "Pointsification" comes as a result of lacking creativity and represents a poor approach to gamification. In order to create a gamified system that truly engages users and affects their needs for satisfaction, a game designer should put more work and effort than just throwing points, badges and leaderboards into the system hoping for users to feel engaged.

The SDT and its sub-theory CET have been explained in Section 2.1 and serve as the foundations on which we base our gamification design. Employing these empirical theories, we aim to reach the goal of positively affecting users needs for satisfaction, which (according to SDT and CET) results in increased intrinsic motivation. Before proceeding with state-of-the-art gamified systems, it is important to understand some key concepts that compose a well designed game. In gamification, the most frequently used framework is the MDA Framework [21]. It is one of the most leveraged frameworks of game design and it is an abbreviation of the terms: Mechanics, Dynamics and Aesthetics.

- **Mechanics** compose the functioning of the game. They allow a designer to have complete control over the levels of the game, giving the ability to guide the players actions.
- **Dynamics** are the interactions of a player with the game mechanics. They determine the action of the player in response to the mechanics of the system.
- **Aesthetics** of the game are the elements that define how the player is feeling during his/her interaction with the game.

By carefully studying and exploring these models and frameworks when working on the game design, we are able to transform a system from being monotone and tedious to something that is interactive and fun to do.

A well designed game or gamified system can be used to gather large amount of data and information. This data can potentially open doors to improving advanced systems such as search engines or helping scientific researchers find solutions to complex protein structures by using the computing power of the people (See Foldit [54]). Entertainment Software Association did a survey to find out what are the characteristics of an average gamer and how much time gamers spent playing video games. The following statistics were reported [2]:

- An approximation of 5 million Americans will spend 40 or more hours a week playing games, which is the equivalent of a full time job,
- 60% of Americans are gamers,
- During 2013 the gaming industry was worth 22 billion dollars, with 16 billion spent on game content only,
- Females account for 50% of gameplay and purchases,
- The average gamer is 36 years old

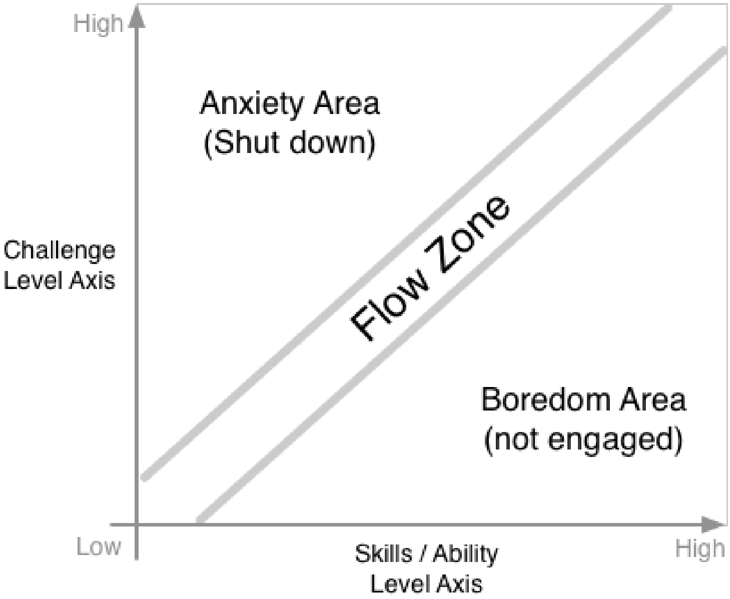


Figure 10: The flow zone of a gameplay [2]

A very important aspect in game design is the ability of maintaining the so called *Flow Zone* during the gameplay. The success of a game is achieved if the player is constantly kept within the flow-zone [2]. As illustrated in Figure 10, achieving flow or being in the *flow-zone* indicates the players' state of not being too overwhelmed with challenges but also avoiding the state of being bored because the game continues to offer easy challenges to the player. Zichermann et al. [53] argues that a game designer must create a careful interplay of the system with the player by relentlessly testing their interactions until the point in which the player is between anxiety and boredom. This rule is applied from the first interaction of the player with the system. This brings us to another quite important point in game design: Onboarding.

According to Zichermann et al. [53], statistics from the casual games market show that the first minutes a player interacts with the game are the most important. This is because during the first minutes the player makes a decision whether he/she likes the game and will continue to play or not. Therefore, onboarding plays a crucial role to the overall success of the game. Onboarding is defined by Zichermann as the act of bringing novice players into the system. The responsibility of this part of the game is to carefully reveal the complexity of the system to the player. In short, the goal of onboarding is to train and engage players but not overwhelm them.

Finally, some aspects which contribute to the idea of keeping the player engaged and motivated during gameplay need to be pointed out before we proceed with the next sections. According to Siemens et al. [2], a player expects the following elements from the game:

- Focused goals
- Challenging tasks
- Clear and compelling standards
- Protection from failure
- Affirmation
- Novelty
- Freedom of choice
- Authenticity
- Affiliation with others

For our Fastype Game we have tried to employ most of these important aspects of game design. As a result, high levels of engagement are achieved and players are intrinsically motivated for playing the game without any other form of incentive except the incentive of being entertained.

4.2 Related Work

The emergence of platforms which contribute to the concept of having a group of individuals commit to creating a collective solution that is far more powerful and robust than individual ones has drawn the curiosity of many industries during the recent years. This concept is referred to as Collaborative Resource Creation (CRC). CRC is being utilized by systems that inherently want to use the creative and powerful nature of human thinking as a source of solving different computationally complex problems. Wikipedia can be considered as the best known example of collaborative resource creation. Furthermore, Open Mind Common Sense (OMCS) movement demonstrated that Web collaboration can be used as a tool to create AI resources. Games can also be considered as CRC system. [15, 22]

Wikipedia and OMCS, as non-gamified systems, rely on users altruism and interest on science in order to commit to contributing. Whereas games provide the feeling of being entertained, and as a result they solely rely on user's intrinsic motivations. Von Ahn et al. [14] argues that the desire to be entertained is a much more powerful incentive than any other incentive. It has been estimated that more than 9 million person-hours are spent by people playing games on the WEB. Dedicating a small amount of those playing hours to contribute to the solution of complex computational problems will result in tremendous benefits. This is the reason why

Games With A Purpose (GWAP) are being frequently used in many domains such as NLP and Semantic Web. Gamifying a system for the purpose of solving or facilitating computationally complex problems can be unquestionably powerful when doing it right. An excellent successful example of such a system is Foldit [54].

Games are also used in education, generally as serious games and as digital game-based learning. In this field, gamification is defined as the use of game-based mechanics, aesthetics and game thinking to engage people, motivate action, promote learning and solve problems [21]. Seaborn et al. [21] argues that gamification has been used as a means of collaborative resource creation by numerous studies which take advantage of the alleged motivational benefits that game design can provide. However, almost all of these attempts of GWAP lack empirical research and standard of practice for design and implementation [21]. In our work, we have attempted to analyze and understand empirical theories such as SDT and CET. They guided the implementation of a GWAP for named entity disambiguation which resulted in a collaborative resource creation system generating training data for supervised NED algorithms.

Seaborn et al. [21] also reported statistics on the usage of gamification across domains ranging from sustainability to health and wellness to education. The findings reported by the study indicate that the fields in which gamification has been mostly applied are Education (35%), health and wellness (13%), online communities and social works (13%), crowdsourcing (13%) and sustainability (10%). They also report that a large majority of applied gamification research did not mention or address any theoretical foundations [21].

Gamification has been used as a mechanism to solve various NLP and semantic web problems. Kaboom [16] is a gamified WSD system that can be classified as a 2D video game in the style of Fruit Ninja game. In this game, players are asked to destroy pictures that are not related to a specific term or concept shown to them beforehand. This approach aims to disambiguate word senses by using pictures as sense descriptors. The pictures that are kept at the end of a game round are considered to be related senses for the ambiguous word. Senses for each ambiguous word in the game were collected manually by expert annotators. Unlike them, we use our implemented microservice framework for generating game data instead of manually creating them, which gives us a head-start in focusing more on the game design aspect. Similar to our findings, Jurgens et al. [16] also reported that game-based annotation systems reduce the cost of producing equivalent resources via crowdsourcing at least by 73% while providing similar quality of annotations. Using Kaboom [16], they also reached a 16.3% improvement in accuracy over state-of-art WSD.

Phrase Detective [15] is another gamified system that was developed to annotate corpora for anaphora resolution. Anaphora resolution is a semantic task used for recognizing that a pronoun like *it* and the definite nominal *the town* refers to some entity as a proper name [15]. They argue that a successful GWAP should make use of all available incentives, namely, personal, social and financial. The game interface should be easy to use, intuitive to learn and designed to engage ones intended player demographic. Furthermore, they emphasize validation as a strong and effective method for quality control. By collecting multiple judgments for each expression, the gamified system can provide quality control and collect useful linguistic data. We employed some of the proposed methods by Phrase Detective as they proved to be appropriate for the design of our game.

To help with the creation of named entities, Green et al. [19] developed the *Entity Discovery* game where players are asked to annotate sentences by marking all named entities found in it. The game is played in pairs. When real players are not online to be paired with, a BOT is used instead. In order to validate the recognized entities by the first game, they developed a second game called *Name that Entity*. The second game was designed as a multiple choice game where the paired players had to choose the type of the recognized entity. The game accepts a specific type for the recognized entity only if the paired players agree upon the type. [19] In contrast, we automated the entity recognition task using our framework and for the validation task we use multiple judgments and rely on agreement levels between players which proved to be very effective and assured high quality of annotations.

Several research studies investigated individual game elements and their impact on intrinsic motivation and performance [22, 55, 56]. Badges, leaderboards and performance graphs, as reported by Sailer et al. [55], positively affected competence and need satisfaction. The same game elements also seemed to contribute to an increase in perceived meaningfulness as it is known that these game elements can create meaning at the game level [55]. Unsurprisingly, Mekler et al. [56] found that *pointsification* (points, levels and leaderboards) functioned as extrinsic incentives effective for promoting performance quantity. They used an image annotation game to determine the effect of these three most commonly employed game elements on needs satisfaction, intrinsic motivation and performance. Using a two-fold experimental study, the game element group performed better in terms of annotation quantity, whereas the quality remained the same. Against their expectations, the different conditions (groups) did not differ in terms of intrinsic motivation or competence as a factor that impacts needs satisfaction. Possible factors that contributed to this outcome is that the game did not provide enough challenge to the players, the feedback was insufficient in determining player performance and the game lacks visual and aural presentation of game elements and feedback to the player. We try to overcome these obstacles by designing a game that keeps the user constantly engaged. Constant engagement in our game comes as a result of increasing challenge as the player progresses, provide visually appealing feedback, empowered social elements and encouraging self-empowering through performance graphs.

4.3 Game Design

Games With A Purpose (GWAP) are implemented in many fields and come in various forms. With regards to their overall game design, they tend to be graphically rich, provide simple interactions and give the player an experience of progression by scoring points, leveling players up and recognizing their effort. Additionally, the design of the game should reinforce quality measures and control the behaviour of players. Encouraging players to concentrate on the task and discourage them from malicious behaviour is crucial for quality control [57]. The goal of our research study was to develop a GWAP that would serve as a tool for validating the named entity disambiguation task. The validation process should be performed at the highest quality level possible and still maintain player engagement by positively affecting their intrinsic motivation: experiencing feelings of competence, autonomy and relatedness while playing. The following subsection explore the different game elements and the corresponding game design principles used in Fastype. It is the design and the appropriate application of these game elements towards solving the NED problem that make this game enjoyable, fun to play and motivate players to come back. This results in generating large-scale annotation data for our named entity disambiguation task without players noticing what the underlying purpose of the game is.

4.3.1 Onboarding

The first impression you get when meeting somebody for the first time is usually a strong indicator that determines whether you will like the company of this new person or not. Similarly, the first impression or the first minutes of interacting with a game are the most important because a player will make a decision whether he or she will continue exploring the game. According to Zichermann et al. [53], the onboarding process is critical to a successful game. A good onboarding process leaves no other options to the player but to win. It is crucial that in this stage the player will be offered with an action at which he will not be able to fail. After having completed the initial action/task, the player should be rewarded for successfully completing it. [53, 22, 53]

During the implementation of Fastype, special attention was put in designing the onboarding phase as effective as possible. Fastype has been designed as a typing game where players have to type on their keyboard as fast as possible to reach high scores and improve their typing skills. Making fast-typing as the central element of our game represents a design decision which has significant impact towards the solution of the problem. By providing paragraphs to be typed by players, we are able to communicate crucial information to the player unconsciously which help the disambiguation process. Additionally, the game requires the user to focus and memorize what they type in their keyboard and also answer quiz-like questions to get additional points, reach new levels and complete various categories. All these concepts have to be carefully revealed to the player during the onboarding stage. According to Zichermann et al. [53], the onboarding process should accomplish the following things:

- Slowly reveal the complexity of the system,
- Positively reinforce players,
- Avoid any action that leads to failure,
- The game should learn something about players in order to personalize the game if necessary.
- All the points mentioned above have to be done within the first few minutes of the player interacting with our game.

Except the points listed above, we also wanted the onboarding process to be slightly intriguing and mysterious in order to tease the player's curiosity. When entering the game site, players are presented with a screen which requires them to know a *game secret* (See Figure 11). Having a secret combination required for entering the game might potentially affect feelings of relatedness. This is accounted by the idea that the player will experience feelings of being part of a secret game club which allows entrance only to players who are aware of the game secret. However, the game secret can be discovered following a specific process. The only way of acquiring it and being granted with access to the game is by following a trial procedure. This trial procedure represents the onboarding process for our game.

The onboarding process starts by asking players to type the words that appear in the screen as small boxes. Points are awarded to the player regardless of their typing speed (rewarding and recognizing the players effort). After the player is familiarized with the way the game works (typing fast while paying attention to the words that appear in the screen), the game continues by introducing a new challenge: a puzzle combined with typing skills. The puzzle consists of several hidden characters that the player has to reveal by typing the words appearing in the screen. The faster the typing the faster the revealing of characters. After the player has revealed all the characters on the puzzle, he is presented with a quiz where the *question* is the revealed word (i.e. the named entity) and the options are the candidate entities extracted from KB. Contextual clues are provided on top

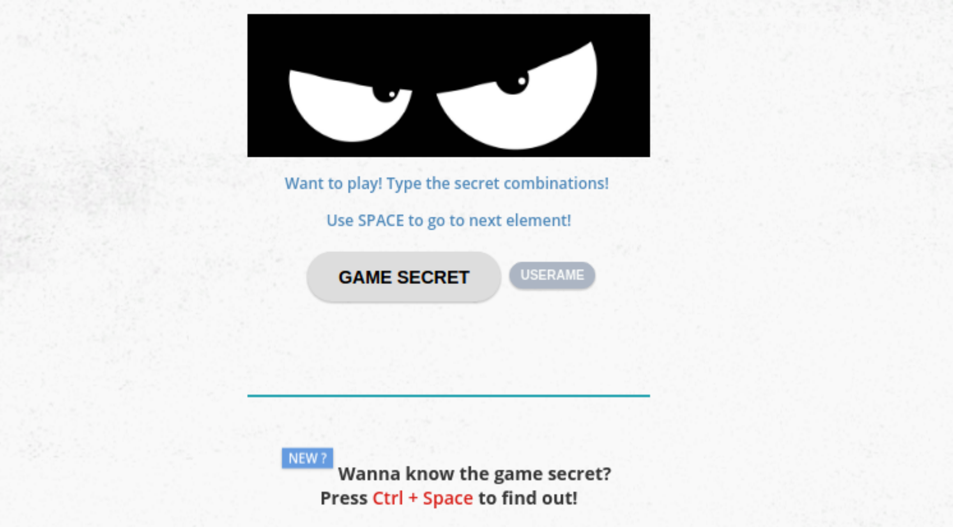

Want to play! Type the secret combinations!
Use SPACE to go to next element!
GAME SECRET
USERAME
NEW ?
Wanna know the game secret?
Press Ctrl + Space to find out!

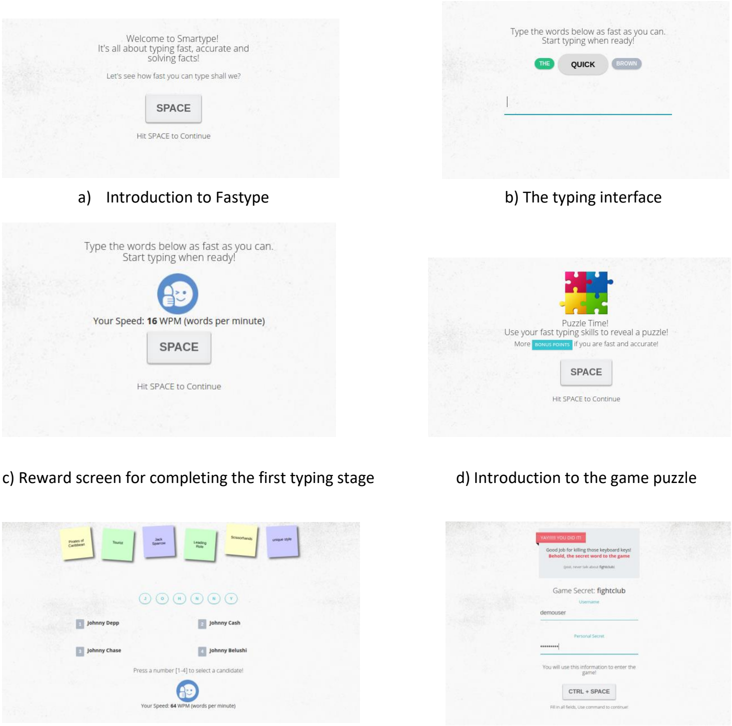
Figure 11: The start-screen presented to the player when entering the game for the first time

of the screen as sticky notes which help the player to make a correct decision. The quiz is a *masked* version of the named entity disambiguation process used in the first experiment. The player proceeds by selecting a candidate option which is rewarded with 10 game points. Please note that if the player selects the wrong option, the system will not proceed until the right option is selected. As a reward for choosing the correct answer, the player finally reaches the end of the trial process where the game secret is finally revealed. Additionally, the player has to register himself with authentication credentials which together with the game secret are used as a secret combination for granting access to the game. Figures 12(a) to 12(f) illustrate the complete onboarding process of Fastype.

4.3.2 Task Design

After exploring and analyzing many gamified systems and the effect of various game elements towards user engagement and motivation, Sailer et al. [55] argues that gamification is not effective per se, but specific game design elements have specific psychological effects. Thus, it is important that gamification is not done just by incorporating a scoring system with some levels to advance and a leaderboard to see your progress. It takes more than that to design a game which attracts and retains a player base.

Chamberlain et al. [22] emphasizes that it is the design of the individual tasks in a gameplay that determine how successful the player can contribute data whilst



Welcome to Smartype!
It's all about typing fast, accurate and solving facts!

Let's see how fast you can type shall we?

a) Introduction to Fastype

Type the words below as fast as you can.
Start typing when ready!

b) The typing interface

Type the words below as fast as you can.
Start typing when ready!

Your Speed: **16** WPM (words per minute)

c) Reward screen for completing the first typing stage

Puzzle Time!
Use your fast typing skills to reveal a puzzle!
More BONUS POINTS if you are fast and accurate!

d) Introduction to the game puzzle

1 Johnny Depp
2 Johnny Cash
3 Johnny Chase
4 Johnny Belushi

Press a number [1-4] to select a candidate!

Your Speed: **64** WPM (words per minute)

e) The game quiz

YAY!!!!! YOU DID IT!

Good Job for killing those keyboard keys!
Behold, the secret word to the game

(psst, never talk about fightclub)

Game Secret: fightclub

Username: demouser

Personal Secret: ••••••••••

You will use this information to enter the game!

f) Final reward of the onboarding stage

Figure 12: The onboarding stage of Fastype

playing. Furthermore, Sailer et al. [55] communicates the importance of game elements being recognized by the player in the gamified environment. Delivering the treatment to the player or to a participant in the game is not enough unless the designer of the game makes sure that the treatment is also received. Failing to design treatments or game elements that are genuinely recognized by the players, results in loss of statistical power and risk to underestimate their effectiveness [55]. In order to adhere to the aforementioned design constraints, significant work effort was placed in designing a task that contributes to generating annotation data of good quality.

A game-round starts by the player selecting a specific category to play or letting the game choose a category in a random fashion for the player. Everything in the game is designed in a way that the player only has to use the keyboard for completing almost every action. The player is then presented with the typing screen where a combination of hidden characters have to be revealed by typing as fast as possible. Figure 13 illustrates the typing screen. A speed radar is placed right underneath the typing text box which displays the current speed of typing. The speed of typing is measured as the amount of words typed per minute while the player is revealing the hidden characters. This element can be considered a strong motivator to encourage the player for typing fast. After having revealed the word, the upcoming challenges/questions all revolve around the text that has been typed during the typing phase. The players are already familiar with this concept since a similar task was already performed during the onboarding phase. Bonus questions are immediately presented to the user after completing the revealing stage (we talk more about the importance of bonus questions towards the overall design in a later subsection). The content of the bonus questions is created by using the text that was typed in the previous stage. Similarly, the most important part of the game that contributes to the original task of disambiguating entities is the quiz. The quiz interface provides contextual clues extracted from our original framework and lists all the potentially candidates (quiz options) from which the player has to choose.

We strongly emphasize the fact that all the game elements presented so far are focused and contribute to the ultimate goal: annotating the entity with the right candidate. The text which the player types during the typing stage, the content of the bonus questions and the contextual clues all represent carefully carved information that contribute in helping the player choose the right candidate. We make sure that the typing text[1] as a game element is recognized and takes the attention of the player. Because if players are able to remember the text they typed, they can

[1] During the fast-typing stage, players are presented with different words that appear on the screen one after another. These words are part of a complete paragraph that is taken from the document where the target entity is part of.

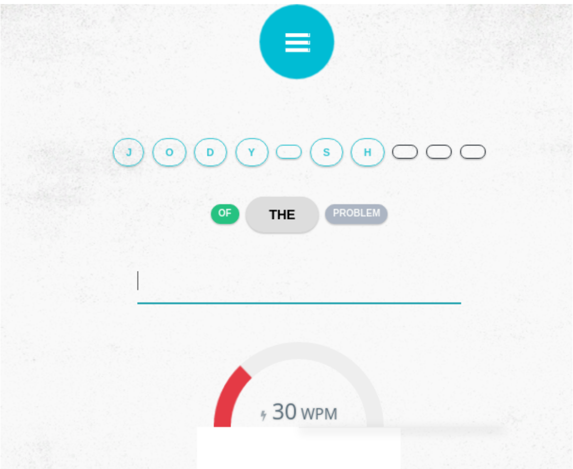
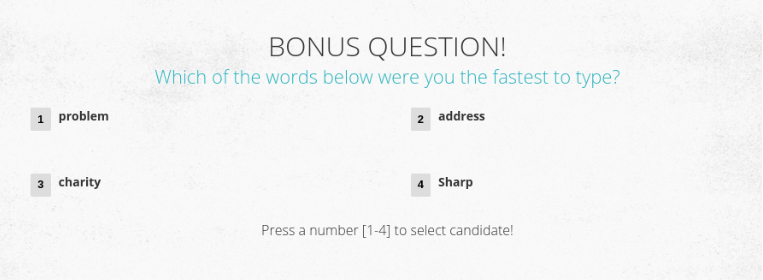


Figure 13: The typing screen and the speedometer calculating the typing speed of the player



BONUS QUESTION!

Which of the words below were you the fastest to type?

1 problem

2 address

3 charity

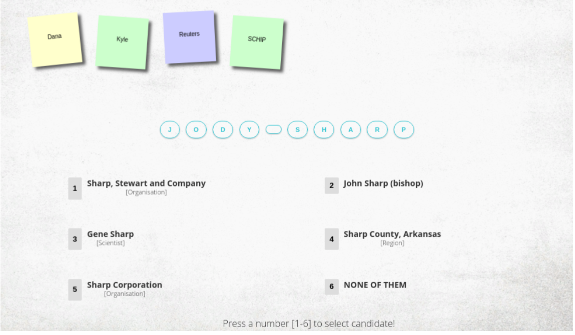
4 Sharp

Press a number [1-4] to select candidate!

Figure 14: An example of a bonus question

get points from correctly answering the bonus questions and also be confident on the candidate option they select. An example of a bonus question is presented in Figure 14 while the quiz interface of the game is illustrated in Figure 15.

Schell [58] (page 214) emphasizes the importance of the skill and chance mechanisms in games. A good game should balance between skill and chance during the gameplay. We believe that we have reached a satisfactory balance between these two factors by equalizing the factor of chance and skill required. The factor of luck is represented by bonus questions which are extracted from the typing text and are genuinely hard to answer since a memorization of everything typed is required. On the other hand, the factor of skills is represented by the quiz element of the gameplay. Correctly answering the quiz requires certain skill-sets from the player in order to maximize their profit point-wise. Except requiring skills from the players, the game should also contribute in allowing the players to master and



1 Sharp, Stewart and Company [Organisation]

2 John Sharp (bishop)

3 Gene Sharp [Scientist]

4 Sharp County, Arkansas [Region]

5 Sharp Corporation [Organisation]

6 NONE OF THEM

Press a number [1-6] to select candidate!

Figure 15: Interface of the game quiz - a gamified version for the named entity disambiguation task

perfection these skills. In general, by playing Fastype, a player will improve the following skills:

- Fast-typing skills,
- Concentration in time pressure,
- Memory training and pattern recognition, and
- Accumulating knowledge in different categories (politics, music, entertainment, health etc).

Improving these skills is one of the goals that players will attempt to achieve. After having observed the players during their gameplay while conducting the second experiment, we are aware that the goals of the game are concrete, achievable and rewarding at the same time. Having all these three elements characterizing the goals of the game is crucial to keeping players engaged and motivated [58].

Reviewed literature suggests a number of criteria for evaluating enjoyment during gameplay. Some of the criteria which apply to task design and which also relate with SDT in terms of their importance towards increasing intrinsic motivation include: Feedback and Immersion [16]. Constant feedback is given to the player for every performed action when interacting with Fastype. Immersion in our case refers to a short and arcade style of the gameplay which make the player experience a deep and yet effortless involvement in the game [16]. A game round in Fastype lasts $4-5$ minutes on average with the player being exposed to different challenging tasks. This results in an effortless and deep involvement in the game for a short period of time. Additionally, the game elements and the feedback provided during the gameplay have been enriched with sounds, visuals and animations. These elements, according to Mekler et al. [56], can positively affect competence, needs satisfaction and subsequently increase intrinsic motivation.

Game elements such as performance graphs, levels, points and leaderboards

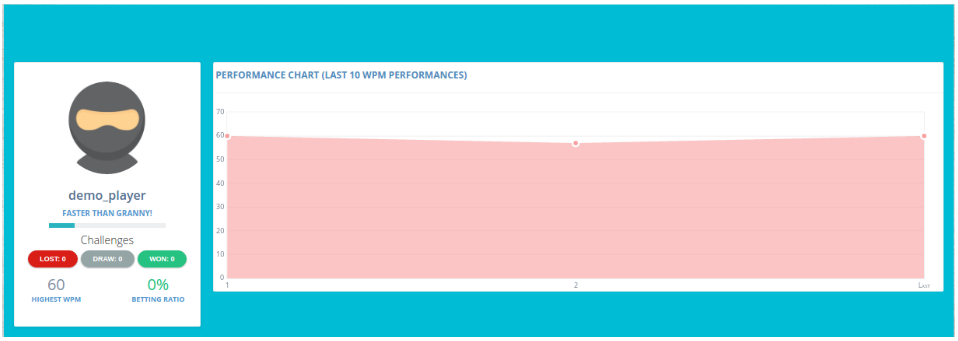


Figure 16: Player profile screen

are provided in the game in order to give the players the feeling of progression and advancement. With these game elements we also reinforce the competitive nature of players to compete with others but also work hard on breaking their personal best scores. This results in encouraging players to get better each time they enter the game (mastery of skills). Figure 16 illustrates the profile of the player where a performance history of the typing speed is provided in addition to level information, challenge status and betting ratio (a balance between the bets lost and won).

The final element that requires our attention is the concept of game-flow. It refers to the idea of keeping the player constantly engaged while not overwhelming him with problems that are too complex to overcome or too easy that the player gets bored quickly. Our idea of keeping the complexity of the game in the same level as the player's progression and skill improvement in the game is by increasing the total number of words that players have to type within a game-round. The complexity is controlled by the level mechanism of the game. When players progress and reach new levels, they are faced with longer and more complex paragraphs to type. However, one can argue that this is not the best and most optimal way of maintaining game complexity and player's flow zone. Being limited in the amount of resources available to be spent in designing and implementing all the game elements, the proposed idea of defining and maintaining game complexity was the only achievable concept within these constraints. However, ideas to improve the existing design of game complexity will be addressed in the discussion section and as such will be accounted for future work.

4.3.3 Freedom of choice

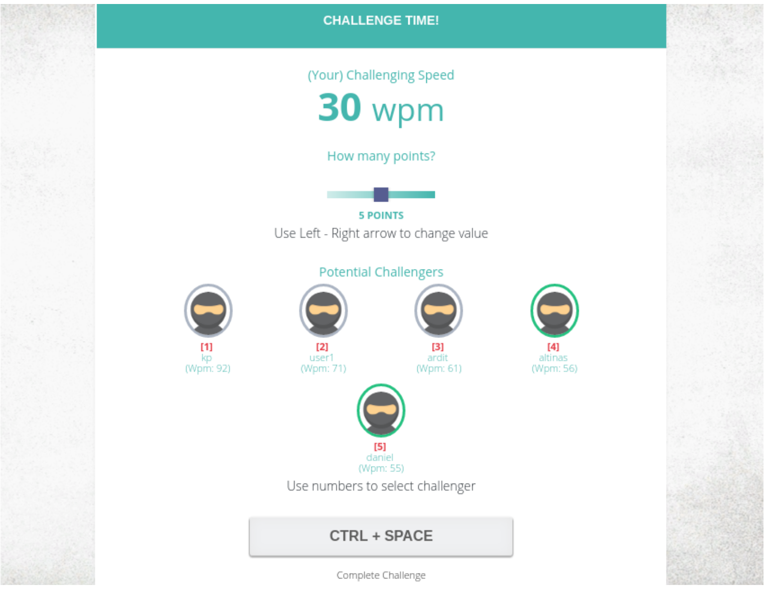
During the first experiment where the non-gamified interface was used to perform annotations, many participants addressed the fact of not having control over the genre of the entities presented to them. Since the selection of the entities was done randomly, one of the participants was constantly getting entities that fell into the Arts category. As a consequence, the participant was feeling very insecure during the task because of being unfamiliar with most of the concepts being presented to him. We would like to stress out the importance of freedom of choice in this aspect. Being able to freely decide what category to play in, is an important factor for reinforcing autonomy and competence. As a result, the game gives the player the freedom of choice by providing several game categories to choose from. For players who like the aspect of surprise and chance, the game can make a random choice for the player if instructed to do so.

Furthermore, the game provides a training phase for players who feel unprepared to take the actual task where the performance is recorded. According to Chamberlain et al. [22] GWAP usually begin with a training phase so that players are able to practice their skills and also show that they have understood the instructions before they do the real task. However, in our case the onboarding stage takes care of explaining the complexity and instructions for performing actions in the game. On the other hand, the training phase allows the user to practice their typing skills. The training phase was also designed for the purpose of getting familiar with a new keyboard since the participants played the game from the experimenters laptop and therefore it was necessary to have a training phase. The game acknowledges the player for the existence of a training possibility in the game by pushing notification on the screen.

4.3.4 Game challenges

Von Ahn in his pioneering work [14] focuses on one type of incentive to motivate players: enjoyment. He emphasized that the main mechanism to make players enjoy a GWAP is by providing them with a challenge. Usually in many gamified systems, challenges are achieved through mechanisms such as requiring a timed response, keeping scores which ensure competition among players, having players with similar level of skill compete against each other and so on. [15]

The design of Fastype strongly reinforces the competitive nature of players by providing challenges which allow players to compete with each other. During each game round, the system keeps track of the player's typing speed, and shows potential challenges based on their WPM (Words Per Minute) accordingly. To assure fair play, the system makes sure that the player is presented with challengers that have roughly the same level of skill. The challenge screen is presented to the player after she has completed the quiz. Figure 17 shows the exact challenge screen which is used by players to send challenge requests to other players in the game. As seen in the figure, players can choose a number of points between 1 and 10 to challenge the



CHALLENGE TIME!

(Your) Challenging Speed

30 wpm

How many points?

5 POINTS

Use Left - Right arrow to change value

Potential Challengers

[1] kp (Wpm: 92)

[2] user1 (Wpm: 71)

[3] ardit (Wpm: 61)

[4] altinas (Wpm: 56)

[5] daniel (Wpm: 55)

Use numbers to select challenger

Figure 17: Fastype Challenge Screen

other players. These points represent the reward or punishment mechanism of the challenge. To assure fair play of the game, players who are challenged will type the exact same words as the challengee. Therefore, an accurate measure of their skill is guaranteed. Finally, having a challenging game where each challenge matches the players skill level contributes to the feeling of competence during gameplay [22].

An additional note that requires attention is the mechanism of punishment in a game. Schell [58] (page 225) addresses the importance of punishment mechanisms in games and if balanced appropriately, they will give more meaning to everything in the game. Punishment mechanisms increase the value of certain game elements such as points. In Fastype, challenges and the betting mechanisms provide a way of punishment for players. In case of failure, their accumulated game points will be taken away, subsequently dragging the player down in lower levels in addition to increasing the failure percentage in their profile. Please note that the players have always the option of skipping challenges or bets on their answers. As a consequence of avoiding challenges and bets in the gameplay, the player will experience a slow progression and less dramatic gameplay as compared to other players who take risks by making bets and challenging others.

4.3.5 Bonus Questions

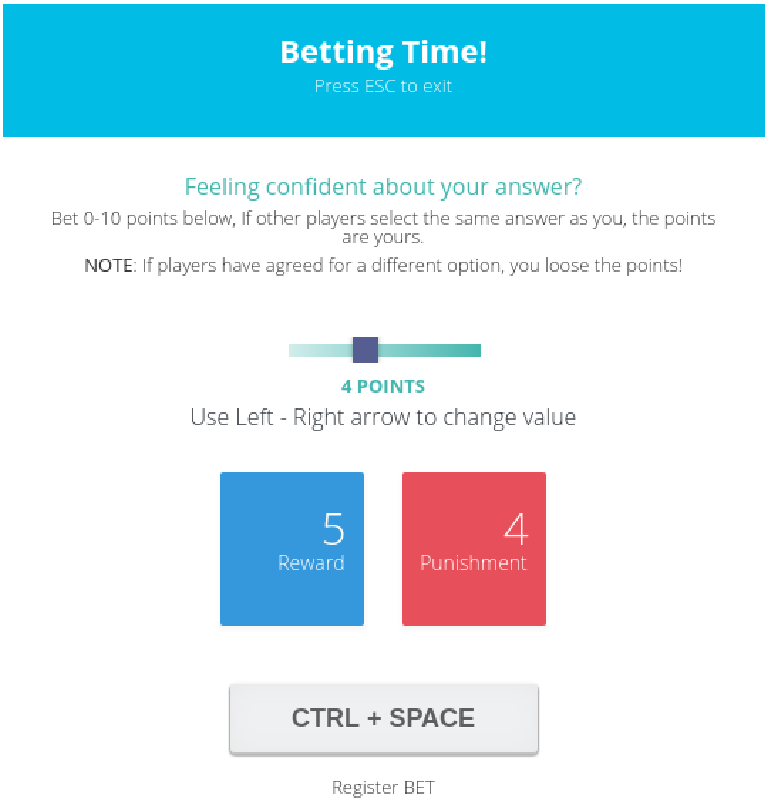
The bonus questions represent the element of chance and surprise in our gameplay. The random nature of the bonus questions give the game a unique flavor of mystery and surprise since the player never knows what to expect from the bonus questions. Therefore, bonus questions encourage players to work on remembering what and how they were typing. Some bonus questions ask about the total number of words typed, the fastest word typed, the number of times a player failed to type a word correctly, the number of times a specific word occurred etc. The bonus question play an important role in the overall playfulness and feelings of enjoyment during gameplay.

4.3.6 Betting System - a measure for quality control

Many of the existing gamified systems point out the importance of quality control in GWAP. Von Ahn et al. [15] suggests two mechanisms for ensuring correctness of data: player testing[2] and repetition or redundancy of data. In our game design we have employed two mechanisms for ensuring that the quality of annotations performed through the game is maintained. The first assessment methodology is accumulating multiple individual answers for a specific entity before deciding whether that answer is correct or not. This corresponds to the redundancy of data for quality control suggested by [15]. The second assessment methodology is using the betting system incorporated into the game. If we look at the betting element in the game from the eyes of a game designer, this element represents the lens of triangularity proposed by Schell [58] (page 212). The lens of triangularity gives a player the choice to play safe for small rewards or take a risk and win big rewards. Triangularity gives an interesting and exciting flavour to our game. On the other hand, if we look at this game element from the eyes of a linguistic expert who demands the generation of trustful and qualitative annotations, this element represents a measure for assessing the confidence of a player towards his choice of candidate. Figure 18 illustrates the betting screen shown to the player immediately after having completed the quiz.

As we can see from Figure 18, the player has the choice of deciding the amount of points he wants to bet on the selected answer. The bigger the number of points used in the bet the higher the risk of loosing or wining them. Players are instructed to think if other players who might potentially encounter the same quiz will also choose the exact same answer as they did. In that case, if they are confident about the given answer, then the players are encouraged to place high bets rather than playing it safe. However, the choice remains completely in the hands of the player. Our betting mechanism works in the way that for one individual entity, when more than 4 unique judgments agree on the same candidate, this entity is considered to be resolved. All the players who's selected candidate is the same as the correct candidate (meaning their answer falls within the absolute majority) will be rewarded

[2] Player testing is evaluating the players output by occasionally matching it against gold standards or already annotated data



Betting Time!

Press ESC to exit

Feeling confident about your answer?

Bet 0-10 points below, If other players select the same answer as you, the points are yours.

NOTE: If players have agreed for a different option, you loose the points!

4 POINTS

Use Left - Right arrow to change value

Figure 18: Betting as a quality control mechanism and representative of triangularity for Fastype

with the amount of points they bet for the particular entity. Similarly, all the players who's answer was not the one selected by the majority are punished by subtracting the amount of points they bet from their overall accumulated points in the game. In terms of annotation quality, candidates with higher betting scores represent higher levels of confidence which means we can be sure that the selected candidate is the correct representative for the target entity.

4.3.7 Social Interaction & Engagement Loop

Significant design and implementation effort was put in making the game as much socially interactive as possible. Among the three SDT elements which have direct impact in intrinsic motivation and needs satisfaction, social interaction contributes to the feeling of relatedness with others. The main social interaction mechanism implemented in the current version of the game are challenges. The desire to compete and overcome players in the leaderboard can be seen as an element that entails social interaction within the game.

The concept of engagement loop on the other hand is a fundamental aspect that needs to be clearly defined in order to maintain player motivation. A successful designed engagement loop always leaves something incomplete or pending in the game for players to come back and play. Game designer David Perry suggests that the key to addictive game design is by creating a game that keeps the player engaged by doing three things all at the time: exercising a skill, taking risks and working out a strategy [13]. Zicherman et al. [53] gives substantial importance to the social engagement loop and defines the following four key aspects to be considered by game designers when thinking about engagement loops: motivating emotion, player re-engagement, social call to action and visible progress/rewards. Table 5 lists the game elements of Fastype that contribute to the different aspects of social engagement loop.

Table 5: Game elements of Fastype that contribute to the social engagement loop

Social Engagement Loop	
Motivating Emotion	Exercising fast typing Competing with others Learning new facts Improving Memory skills
Player Renegagement	Levels Challenges Increasing WPM Completing Categories
Social call to action	Challenges
Visible progress	Leaderboard Points Levels Challenge Win/Lose Ratio

4.4 Methodology

The last proposed research question posed by this study is concerned in finding out what game mechanics and design principles are best fit for our research problem in order to positively affect intrinsic motivation and achieve high levels of engagement. We are also concerned to find out whether the implemented game contributes to qualitative annotations by non-experts. Being able to appropriately use the knowledge of non-expert annotators, we can harvest the potential of players with different levels of expertise in annotating, age group and academic background. To be able to answer our third research question and assess the usability and performance of our implemented game, we conducted another experimental user study with the same participants who participated in the first experiment. However, a different dataset was used for the second experiment. Choosing a different dataset was important since the participants were already familiar with the two datasets used in the first study.

MSNBC first introduced by [28] is the dataset used for our second user study. It contains news-wire text from MSNBC news network. The dataset was created in 2007 and contains information and facts from that period. This fact was communicated and acknowledged by the participants on the second experiment in order to avoid reference errors such as linking the entity "USA President" with the candidate "Donald Trump" instead of "George W. Bush". The complete dataset contains 20 documents in total. However, we used only 12 documents in order to reduce the total number of entities to deal with. Experimenting on a reduced version of the dataset instead of the complete version does not have any implications on our results. The reason we used a reduced version of the dataset is because our game relies on multiple judgments to disambiguate an entity (4 independent judgments to be precise). Therefore, having a small pool of entities to pick from would result in more entities being resolved during our time limited user experiment. The more entities resolved during the experiment the more confident the results of our analysis. In summary, we used the documents in the following categories from the MSNBC dataset for our second experiment: Entertainment, Health, Politics, Technology, Sports and World. The gold standard for MSNBC reported 251 named entities in total with an average of 20 named entities per document.

Similar to the first experiment, emails and social media were used as communication sources for recruiting participants for the second experiment. Initially, we aimed to recruit the same participants who took part in the first experiment, however, since some of the participants reported to be away or sick during the time of experiment, we sent invitations to others as well. As a result, 26 participants took part in the second experiment with 24 having participated for the second time.

The university campus was the environment where the second experiment was conducted. Participation was scheduled in different time-slots with a maximum duration of 40 minutes assigned for each round. Stretching the session duration to 40 minutes has been done for the only reason of giving temporal space for participants who wished to play longer if necessary. During the invitation process, it was made clear to the participants that the (mandatory) duration of the experiment would be approximately 20 to 25 minutes including the pre-questionnaire, gameplay and post-questionnaire. Similar to the first experiment, we used a consent form to communicate the purpose and the voluntary nature of the experiment. In order to assure consistency of data given by participants, a pre-questionnaire gathering demographic information was also used in the second experiment in addition to questions related to frequency of playing video games. We report on the demographic data of participants and their frequency of playing video games in section 4.5 whereas the questions used in the pre-questionnaire are presented in Appendix B.

The completion of the pre-questionnaire was followed with an immediate transition to the gameplay. In order to be consistent with the first experiment, players were asked to perform game rounds iteratively until the experimenter instructed them to stop. 15 minutes were dedicated for performing game rounds (annotations). However, the game includes an onboarding phase which introduces the player with the game and gradually reveals the complexity without overwhelming the player with too much information at once. Therefore, the time used by each participant to complete the onboarding phase is not recorded as part of the 15 minutes of total gameplay. The reason is that no annotation is performed during the onboarding phase of the game. After the 15 minutes of gameplay, participant were told to either finish playing or continue as they desire. This methodology was used in the first experiment as well and allows us to assess the participant's engagement with the game and its fun aspect. Finally, a post-questionnaire was used to assess player engagement, motivation, usability and enjoyment of the game. We report on the results of the post-questionnaire in section 4.5 and list the questions used in this questionnaire in Appendix B.

Agreement level with the majority and comparison of game annotations with the gold standard are the assessment methodologies used for determining the quality of annotations performed by players while playing Fastype.

Additionally, in order to report confident results, it was necessary to perform ANOVA analysis between the two conditions (AnnotateMe and Fastype) for different factors such as: enjoyment, level of engagement and the likeliness of participants to use each interface for an actual task outside the experiment. For the sake of reproducability of results, it is important to note that the observations used to calculate ANOVA for the two conditions were not the same[3]. The observations from the second experiment were slightly smaller than those from the first experiment. However, from a statistical point of view, small differences between conditions do not affect the final results when calculating one way ANOVA. Please note that ANOVA analysis were manually performed using Excel spreadsheets. Finally, a third assessment questionnaire was sent to all participants who took part in both experiments in order to assure that the game was significantly more engaging as compared to the plain interface. The questionnaire consisted of one question which asked about their preferred interface for performing the task of resolving named entities. The content of the questionnaire is presented in Appendix B.

[3] 31 Participants for the first experiment and 27 for the second experiment

Hypothesis

The second experiment was conducted in order to determine whether our proposed gamified system can intrinsically motivate users to perform a tedious and boring tasks such as named entity disambiguation. Another important aspect that needs to be emphasized is the quality of annotations. It is crucial for the design of the game to avoid having distracting elements that would intentionally or unintentionally degrade the quality of annotations. Keeping these important aspects in mind, we hypothesize that:

- H2.1: By employing game mechanics and design principles based on empirical research and game theory, players will be intrinsically motivated to play the game!
- H2.2: In case H2.1 is supported, we can further hypothesize that users who have used both interfaces to perform annotations will choose the game significantly more than the non-gamified interface!
- H2.3: The quality of annotations performed by players will not be degraded even after implementing game elements that do not directly contribute to the annotation task!

4.5 Results

This section reports on the results after conducting the second experiment and analyzing the effects of the different game elements towards user engagement with Fastype. We also report on the playfulness of the game, feelings of competence, autonomy and relatedness in addition to testing and comparing the quality of annotations with the first experiment. In overall, these analysis will guide the rest of this research study towards answering the last research question regarding gamification and the corresponding hypothesis.

4.5.1 Participants

It was mentioned in the previous section that the recruitment of participants has been done in a consistent way with the first experiment. The participation resulted in having 27 participants in total, 3 participants less than the first experiment. Those who did not show up for the second experiment were either not at the university campus during the whole week of experiment or were sick and could not show up. However, among those 27 participants 11.1% (namely 3 out of 27) of them were participating for the first time. This small group of participants that were not part of the first user study were not presented with questions that involved comparisons or assessment of AnnotateMe with Fastype.

The pre-questionnaire for the second experiment was designed for gathering information about players' demographic data in addition to information related to their gameplay frequency. The questions related to gameplay frequency will provide approximate insights on how often participants play video games and see if this is a potential bias on the quality of annotations and evaluation of the game. Regarding demographic data (since 89% of participants were participating for the second time) we report similar results: the average player is a 25 years old male student who studies in a computer science related field and occasionally plays video games with an average of 8 hours played during the last month. Figure 19 presents two graphs which report on the gameplay frequency (left) and average hours played (right) during last month. From the results of the pre-questionnaire, it can be concluded that our participants in general can not be considered as expert gamers based on their frequency of gameplay. However, on average the participants have a mix of different gaming frequency as observed on the left graph in Figure 19. Therefore we argue that the population sample used for this experiment represents the potential average player of Fastype outside the experiment.

4.5.2 Annotation Performance

All the data used in the game has been automatically generated by the framework. The only task which was manually performed by the experimenter was uploading the MSNBC documents into the framework through the admin interface. The

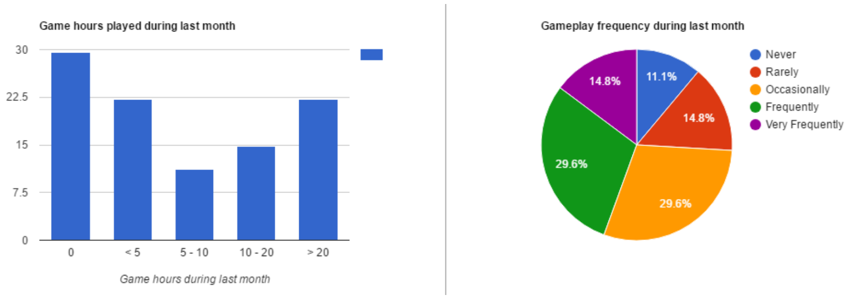


Figure 19: Results on participant's gameplay frequency for Experiment 2

rest was carried out automatically from recognizing the entities to generating the corresponding Dbpedia candidates.

Regarding the number of entities available in the MSNBC dataset, our NER microservice recognized 228 entity mentions in total. Corresponding F-Measures were used to calculate the performance of the entity recognition service using MSNBC dataset and resulted with 0.77 for precision, 0.83 for recall and 0.8 for the f-score. The performance of other automatic annotators in terms of entity recognition are presented in Table 6. We observe a very small improvement on the overall F-score of our framework compared to AIDA which performed best among the other automatic annotator. However, improving entity recognition performance is not the scope of this study and therefore we will not elaborate on the respective NER performance results.

Table 6: NER performance on MSNBC dataset

-/Dataset	**MSNBC Dataset**		
Annotator/Metric	**Precision**	**Recall**	**F-Score**
Babelfy	0.45	0.65	0.52
Dbpedia Spotlight	0.58	0.6	0.57
AIDA	0.92	0.71	0.78
Dexter	0.49	0.66	0.39
Fastype (Experiment 2)	**0.77**	**0.83**	**0.8**

The performance of the automatic annotator used by our framework, namely Dbpedia Spotlight, in terms of accuracy of linking the identified entities with the correct candidate has been calculated as well. When querying Spotlight for a specific entity, a list of candidates is returned correspondingly with each candidate associated with a final score. The candidate with the highest final score is considered by Spotlight to be the correct link associated with the target entity. We run our analysis on all the 228 recognized entities and checked for each entity whether the associated candidate with the highest final score was the correct representative. Evaluation was carried out based on the gold standard of MSNBC provided by GERBIL [52]. Among the 228 recognized entities, Spotlight linked 66.35% of them correctly and 33.65% incorrectly. This shows a significant improvement compared to the datasets used in the first experiment. A possible explanation for the observed improvement is the nature of the dataset. The entities recognized in MSNBC have a less ambiguous nature compared to the KORE50 and Spotlight datasets used in the first experiment.

Before proceeding with reporting on the performances of the non-expert annotators, we report on the total number of entities resolved during the experiment. The assessment technique used for evaluating an entity based on the participants judgment is the same as for the first experiment. The game accumulates four different judgments from independent players before resolving an entity with a specific candidate. Among the 228 entities recognized in total, only 24 were resolved. Unfortunately, this is 60% less entities resolved in the second experiment compared to the first one. The reason for this outcome is because the average tasks/game-rounds completed during a session dropped significantly for the game. The game elements such as fast typing, bonus questions, bets and challenges are all additional tasks that take significant amount of time to complete. Besides the fact that the game elements enrich the experience, improve player engagement and increase intrinsic motivation, they did slow down the annotation process significantly more compared to AnnotateMe Interface. One might argue that drawing conclusions on such a small sample of resolved entities cannot be generalized. We recognize this issue in our analysis and therefore we address it as a limitation on our results. However, the potential of the game is significantly higher compared to the non-gamified interface used in the first experiment. Therefore, we predict that the game will still perform much better in terms of user engagement and maintain annotation quality regardless of the number of annotations resolved.

Among the 26 resolved entities 95% were correctly linked with a Dbpedia candidate whereas 5% were incorrectly linked when compared to the gold standard. Table 7 shows the performance of the game and Dbpedia Spotlight on all 26 resolved entities in terms of precision, recall and f-score measures. The game which used non-expert annotators exhibits a slight improvement compared to the automatic annotator. However, after running a one-way ANOVA, the differences between the two groups are not statistically significant for p = 0.05. Regardless, for the entities that have been resolved, the annotation quality remained unchanged as compared to the first experiment, with a very small improvement for the second experiment.

With the results acquired so far, the third hypothesis (H2.3) for the second experiment is supported, meaning that the quality of annotations was not degraded as a result of gamifying the NED system.

Table 7: Annotation Performance - Experiment 2

Precision, Recall, F-Scores				
Annotator	**Precision**	**Recall**	**F-Score**	**DATASET**
Game	**0.87**	**0.95**	**0.91**	**MSNBC Dataset**
Dbpedia Spotlight	0.85	0.81	0.8	MSNBC Dataset

In the first experiment we reported on the agreement levels of participants which represents the ratio of correct answers given compared to the total number of answers. Even though the annotation quality remained the same, the agreement level experienced a significant drop during the second experiment. Figure 20 presents the agreement levels of all participants for the second experiment. The average agreement level for an individual participants for the second experiment is 32% which is significantly less than compared with the first experiment which was 51%. Please note that the agreement levels do not take into consideration annotation data for entities that have not yet been resolved. For example if three participants have selected the same candidate for a specific entity, this means that they agree on that specific candidate. However, unless a fourth participant selects the same candidate, the entity is not resolved and therefore the other three participants are not credited for having agreed with each other. We argue that the performance drop is a result of having 60% less data for the second experiment to analyze and provide results as compared to the first experiment.

4.5.3 Game Design Analysis

In the previous sub-sections we reported on the performances of non-expert human annotators in terms of annotation quality and agreement level and saw that the quality of annotations did not experience any decrease. In this subsection we report on the performance of Fastype in terms of playfulness, player engagement and see if the players were intrinsically motivated to participate and play the game.

One of the metrics used to assess the playfulness or the attractiveness of both interfaces (AnnotateMe and Fastype) was by measuring the number of the so-called *free-will annotations*. When conducting each experiment, the participants were instructed to perform their tasks continuously until instructed to stop. An annotation was considered to be a *free-will annotation* when participants free-willingly decided to continue performing tasks even after the experimenter informed them that they

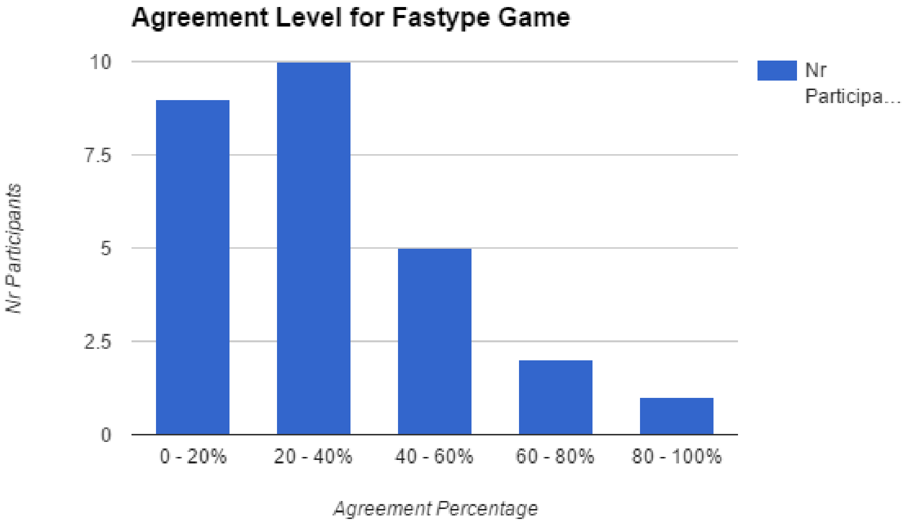


Figure 20: Participant's Agreement Level for Experiment 2

have already completed the number of obligatory annotations and were allowed to finish the experiment. Table 8 shows the results of free-will annotations performed in both experiments in addition to the average number of tasks performed in a session. The game exhibits a slight improvement over the non-gamified interface in terms of free-will annotations. However, the calculation of one-way ANOVA resulted in statistically insignificant difference for p = 0.05. The reason for the insignificant difference can be explained by observing the second row of Table 8, namely, the average annotations performed during an experiment session. We observe that participants performed 50% more annotations using the non-gamified interface as opposed to the game. Besides the fact that the difference was not significant, we argue that an average of 23% free-will annotations over an average of 10 game rounds performed by a single participants is better than an average of 17% free-will annotations over an average of 22 annotation rounds. In a long run, we are confident that the game would perform significantly better and have a significant higher attractiveness as compared to the non-gamified interface. The results presented in Figure 21 easily back up this claim.

After both experiments were conducted, we asked participants that took part in both experiments to fill-in a final questionnaire regarding their preferred interface in case they would be asked to do a third experiment. As seen from Figure 21, the majority of participants preferred the game compared to AnnotateMe Interface. Please note that the participants had the possibility to choose a third option which

Table 8: Results of free-will annotations for experiment 1 and 2

Participant Average Freewill and total annotations performed		
	Fastype Game	**AnnotateMe Interface**
Average of free-will annotations	23%	17%
Average annotations performed	10	22

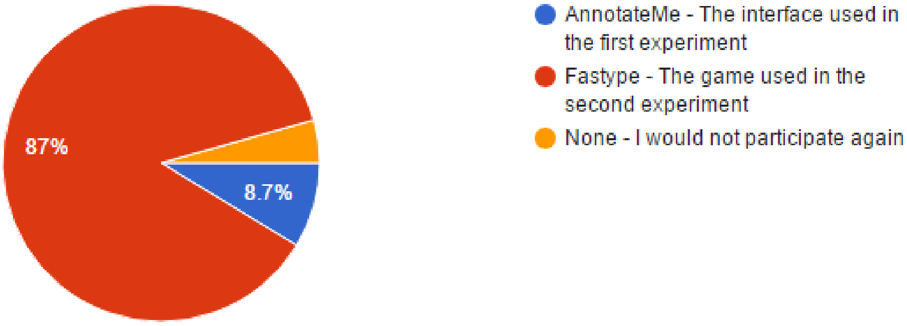


Figure 21: Participant's preferred interface for performing annotations

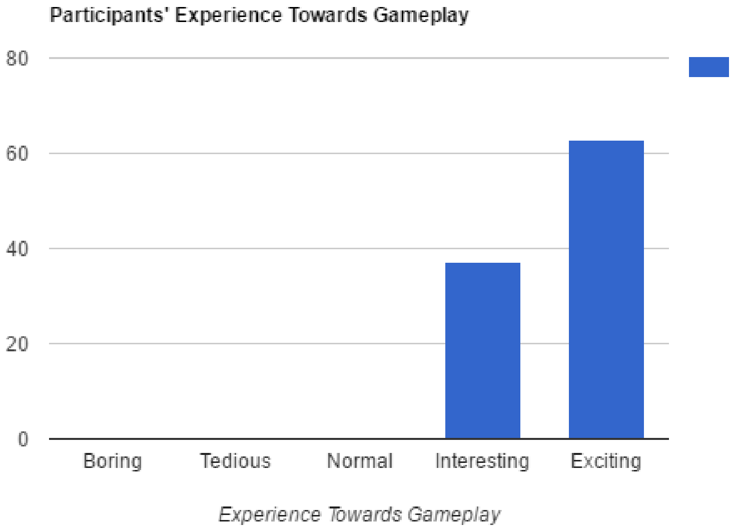
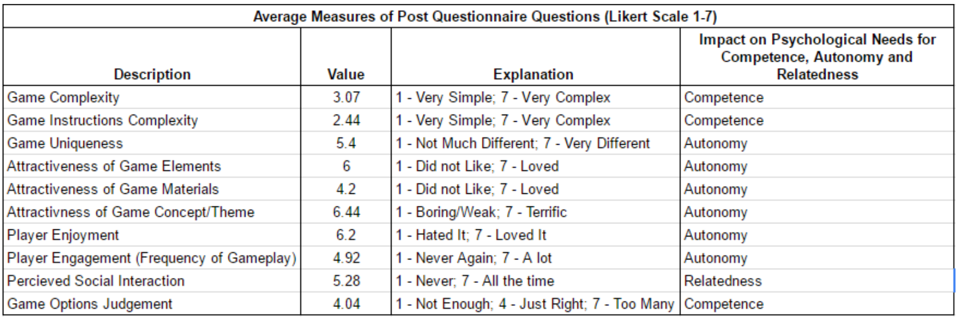


Figure 22: Participant's experience towards the gameplay

allowed participants to express feelings of neutrality or dislike for both interfaces. The third option was provided as a measure against potential bias in our results by allowing all participants who disliked both experience to freely express it. The questionnaire was completely anonymous and was sent through emails where participants completed it from home.

A post-questionnaire assessing the overall design of the game was also used in the second experiment. In the post-questionnaire, participants were asked different type of questions with each contributing to the assessment of different aspects of game design. Additionally, in order to compare the attractiveness and engagement of participants with the game and the non-gamified interface, both designs were assessed using similar questions in both post-questionnaires.

In order to assess the attractiveness of each interface we asked participants to rate their experience with each interface. The results of the first experiment are shown in the top-right graph presented in Figure 8 whereas the results of the second experiment are presented in Figure 22. To make sure that the game was perceived as significantly more attractive and fun to use as compared to the non-gamified interface, we run one-way ANOVA analysis. Results confirm that the difference is statistically significant for $p = 0.05$ and for $p = 0.01$. This indicates that the game was perceived significantly more attractive as opposed to the non-gamified interface.



Average Measures of Post Questionnaire Questions (Likert Scale 1-7)			
Description	Value	Explanation	Impact on Psychological Needs for Competence, Autonomy and Relatedness
Game Complexity	3.07	1 - Very Simple; 7 - Very Complex	Competence
Game Instructions Complexity	2.44	1 - Very Simple; 7 - Very Complex	Competence
Game Uniqueness	5.4	1 - Not Much Different; 7 - Very Different	Autonomy
Attractiveness of Game Elements	6	1 - Did not Like; 7 - Loved	Autonomy
Attractiveness of Game Materials	4.2	1 - Did not Like; 7 - Loved	Autonomy
Attractivness of Game Concept/Theme	6.44	1 - Boring/Weak; 7 - Terrific	Autonomy
Player Enjoyment	6.2	1 - Hated It; 7 - Loved It	Autonomy
Player Engagement (Frequency of Gameplay)	4.92	1 - Never Again; 7 - A lot	Autonomy
Percieved Social Interaction	5.28	1 - Never; 7 - All the time	Relatedness
Game Options Judgement	4.04	1 - Not Enough; 4 - Just Right; 7 - Too Many	Competence

Figure 23: Results from the post-questionnaire analysis for experiment 2

Regarding the player's perceived engagement with both interfaces, we asked participants how often they would use each interface outside the experiment. Please note that for the same question we unfortunately used different type of answers. The answer for the first experiment corresponded of labels ranging from "Definitely Not" to "Definitely", and the answer for the second experiment corresponded a 7-point liker scale ranging from "Never Again" to "A Lot". However, when calculating one-way ANOVA, we normalized the 7-point liker scale to 5-point liker scale (dividing the value with 7 and than multiplying it with 5) to be able to properly compare the two observations together (See Appendix C.2). The results of the one-way ANOVA report a statistically significant difference between the two observation for $p = 0.05$ as well as $p = 0.01$. These results indicate that the game was perceived as significantly more engaging than the non-gamified interface.

Finally, Figure 23 presents a table with the final results for the rest of the questions used to assess the design of the game on the post-questionnaire. It can be observed that the game scores relatively well in all of the design questions presented to the participants. The table also illustrates the impact of the different game design elements on basic psychological needs for autonomy, competence and relatedness which represent the main factors for affecting player's intrinsic motivation. Since the players perceived the game as generally enjoyable, engaging, socially intractable, easy to adapt with the instruction and rules of the game, genuinely liked the elements, materials and the theme of the game, we claim that the participants were intrinsically motivated to play and do not rely in any other incentive but the desire to be entertained. Consequently, based on these acquired results the first hypothesis of the second experiment is supported (H2.1). The second hypothesis (H2.2) is also supported since the one-way ANOVA analysis indicate that the participants preferred to use the game significantly more than using the non-gamified interface.

5 Discussions

Named entity linking and disambiguation has been a problem for which many studies have tried to find an effective solution. An appropriate solution to this problem can be either by using automatic supervised algorithms or by utilizing human input for validation and generation of annotation data [43, 39, 10, 20, 44]. Despite the advancement of supervised techniques using machine learning algorithms, tasks such as NED and WSD still have not reached a satisfactory performance for generating trustful quality of annotations [4]. Relying on human input for validating these automatic approaches has been seen as the only way for progressing in this aspect. A question which has drawn constant attention is whether non-expert human annotators are capable of generating annotations with a quality comparable to expert annotations. The implementation of a complete NED framework which was used to generate annotations using non-expert annotators in an experimental setup reveal that these users are capable of generating annotations with an accuracy of 0.92 (F-score).

5.1 A framework that supports qualitative annotations

Our findings are complementary with previous work which also tested the ability of non-expert users to perform tasks for NED and WSD [59, 20, 39, 45]. However, previous human centered approaches for NED were either focused on assessing the usability of their interfaces [13, 20, 60] or they used human annotators for temporarily validating automatic generated data [44, 45, 46]. Supervised and semi-supervised approaches on the other hand have been implemented as complete frameworks without relying on human evaluation of annotation data [43, 42]. Our study however, provides a solution which utilizes best systems and techniques from the automatic approaches and usability best practices from human centered studies. It refines on previous work by effectively composing such combination into a singular micro-service framework. It is a framework that takes the best out of automatic extraction algorithms while harvesting human knowledge on essential parts such as validation and disambiguation. Analysis and findings from the results acquired during experimental trials throughout this research project show that it is possible to rely on non-expert human annotators for generating qualitative annotation data. This data can be used to enrich HTML content with semantic information from LOD knowledge bases or train supervised approaches until they become mature enough to disregard human judgments.

Additionally, an important point that deserves attention and ought to be discussed is the assessment and comparison methodology used to evaluate the performance of the framework that uses the target population for generating annotations. As pointed out in the methodology section 3.4, one might argue that comparing annotation performance of human annotators with supervised algorithms might not be a valid or fair comparison. We would like to stress out and emphasize the fact that this research study is concerned in using non-expert annotators for generating large-scale annotation data which can be used for improving supervised algorithms for NED. The significant small number of available expert annotators, and also the extensive amount of work required by such experts to generate annotation data on large scales, makes it impossible to rely on this rather condensed user base. Our ultimate goal is to advance supervised machine learning algorithms for NED or WSD. In order to achieve that, the training data which is used to feed these algorithms must be of high quality since the performance of these algorithms is dependent on the quality of training data. Having a model (in our case the gamified framework) that assures the generation of large-scale and qualitative annotation data represents a huge step towards advancing the field. Therefore, we argue that our decisions in these respective assessment and comparison methodologies are appropriate for the direction on which this research study was focused. However, there are additional direction which might be interesting to follow in future work. It would be interesting to test the annotation performance of non-expert annotators with those of expert annotators. Though, we believe that an annotation quality of (f-score) 0.92 for KORE50 and Spotlight dataset and 0.91 for MSNBC dataset represents roughly expert level of quality generated by non-expert annotators, a future research study in this direction would improve the validity of our claims.

5.2 Short clues as context representatives

During the time of the writing, research studies that were focused entirely or partially on formulating and presenting contextual information to human annotators for performing disambiguation were non-existent. However, many studies were focused in defining surrounding context of ambiguous entities as features and using them as additional attributes to supervised and semi-supervised algorithms [28, 26, 46]. Based on the results reported by these studies, providing contextual information as features improved the performance of such algorithms. However, these features are designed for supervised algorithms and may not be as helpful for human annotators.

This study extends on previous work by providing a service that can effectively generate contextual information surrounding the target entity that might potentially help non-expert annotators easily disambiguate ambiguous entities. Our findings reveal that participants tend to prefer short contextual clues instead of complete sentences. However, limitations in the methodology chosen to assess the effectiveness of these clues diminish the confidence of our claims. Partially confident, we claim that the clues generated by our service provide sufficient information for annotators to correctly disambiguate an entity. The main reason that resulted in choosing a rather poor and inappropriate evaluation methodology for assessing the effectiveness of contextual clues is the lack of time resources in conducting a third experiment. Asking participants in a post questionnaire whether the context clues were useful or not without testing such variable in a two-fold experimental study is not the most appropriate assessment methodology. We acknowledge such fact as a limitation to our study and propose a different assessment methodology that should be used in future work to validate our claims.

The technique used in the game for communicating the context to the human annotator deserves attention at this point. In addition to the short contextual clues, players were exposed to the complete sentence. The way in which the sentence was exposed to the players did not exhibit frustration or extensive cognitive load. The design of the game task made it possible to use the sentence for communicating the context unconsciously by having the user type the complete sentence as part of a fun activity (fast typing). This proved to be significantly effective since the quality of annotation performed in the game increased by 7% as compared to annotations performed with AnnotateMe Interface.

5.3 Gamifying non-gaming systems

With results supporting the claims of generating qualitative annotation by non-expert users using our framework, the gamification process was applied in order to intrinsically motivate users for contributing with annotations [39, 45, 46]. In general, our findings corroborate with previous work in gamification that game mechanics and game design principles can be applied in non-gaming context in order to make the task more engaging and fun to interact [15, 55, 19]. Previous studies reported that gamification of WSD helped in improving annotation quantity, but were not able to acquire significant results to report on quality improvements [56]. Our results extend on this previous work that gamification can be used for not only increasing annotation quantity but also maintain high levels of quality. Analysis and results of our study continue to support the idea that applying game elements in non-gaming context does not necessarily motivate users, improve quality and quantity of annotations. Psychological needs for satisfaction have to be positively affected in order to increase intrinsic motivation and as a result improving quality and quantity of annotations.

Therefore, when designing the individual game elements and constructing the tasks composing Fastype, the three different factors for psychological needs for satisfaction where kept in mind. These three factors included competence, autonomy and relatedness. Referring to SDT theory, positively affecting these factors results in subsequently affecting intrinsic motivation. Based on the results acquired, we discuss the various elements of our game with regard to affecting intrinsic motivation. We argue that the onboarding process and the other game instructions used in the game contributed in positively affecting needs for competence since participants rated the complexity of the game on average 3.07 out of 7. Additionally, the number of available choices provided in the game, were judged as appropriate by the participants. Thus, it is another factor that also affects competence. Psychological needs for autonomy were positively affected by a number of elements composing our game. As illustrated in Figure 23, all the game elements which affected autonomy were positively rated by all participants. We observe that participants perceived the game to be unique with attractive game elements and material, interesting theme (game concept) and significantly enjoyed playing the game (6.2/7). Finally, the psychological need for relatedness was positively affected since participants perceived the game to be socially interactive with a score of 5.28 out of 7 (See Figure 23). Having analyzed and discussed these aspects of the game, we are confident about the fact that the game was perceived as engaging and enjoyable and therefore players were intrinsically motivated to contribute.

A well designed game makes sure its player base is maintained by continuously engaging the players and motivating them to come back without hesitation. Therefore, it is important to address which elements of the game account for retaining players in long periods. Fastype is designed to engage players and motivate them to exercise certain skills and using these skills for performing disambiguation of named entities. During the gameplay, the player is always exercising fast typing, taking risks by challenging other players and betting on the confidence of their answers. These two factors account for addictive game design as suggested by Perry [13]. We argue that challenges, game progress, skill mastery and the desire to be a fast and unbeatable player in Fastype, are the prominent factors which keep the players motivated to come back and interact with our game.

5.4 Limitations

Results from the second experiment reveal that the game mechanics and the overall design of the game positively affected users needs satisfaction and intrinsic motivation to contribute. However, it is important to address that there might be some potential bias in assessing the users initial motivation to participate in the study. In our experiment, we recruited participants that were mainly bachelor or master degree students studying at a higher education institute. It is known from previous research studies that participants from the university usually engage voluntarily in studies. As a consequence, it is possible that they already had a minimum level of intrinsic motivation from the get-go effect, which might have affected the results [56]. Additionally, the population used in the experiment consisted of students with higher education in the field of computer science, information security and interaction design. Thus, one might argue that the label *non-expert annotators* cannot be applied to this type of population. However, regardless of the population sample used in the experiment, the design of the game assures that every player goes through the onboarding process. As a gamified version of a training stage, onboarding assures that players understand the game elements and the concept of the game. Consequently, the game will prepare all players to perform qualitative annotations regardless of their educational background or level of expertise in text processing. Therefore, we stay behind our claims that the framework and the game design supports the generation of qualitative annotations by non-expert users.

To further strengthen our claims in this regard, more research is required to investigate how users' initial motivation to engage in gamified application affects their subsequent motivation [56]. Additionally, testing the game with a larger and broader player base that more accurately represents the general target population would potentially sustain our research claims.

5.5 Practical and theoretical implications

From conducting this research study, we have learned that game design can be applied to non-gaming contexts as long as the design conforms to empirical and psychological foundations. Additionally, in order to have a GWAP that entertains players and has a strong focus on the main task (disambiguation of entities), all game elements should directly or indirectly contribute to the solution of the problem being addressed by the game.

The real challenge in designing a game with a purpose is finding a model that is appropriate for the task but also contributes to user engagement and intrinsic motivation. The main design elements, when addressed appropriately, that contribute a great deal towards having an enjoyable GWAP include:

- controlling players for malicious behaviour,
- maintaining players within the defined flow-zone of game complexity,
- providing meaningful content, and
- non-formal feedback that is visually and aurally attractive without shifting the players concentration and focus from the main task.

In Fastype, all the game elements contribute to the general idea of providing qualitative annotations and each game element plays an important role in helping players make correct disambiguation decisions. Encouraging competition for competitive players without degrading the experience for other players that prefer a more play-safe gameplay are design factors that characterize a well-thought GWAP. Although, some research studies which have investigated on gamifing WSD and entity linking argue that text-based games are limited in their potential compared to 2D video-games, our results conflict with these claims [61]. This research work, based on statistically significant results, rejects the claims stated by Vannella et al. [50]. Analysis of our acquired results prove that even text-based GWAP can be as engaging and entertaining as 2D video-games. Therefore, it ultimately comes down to the design of the game complementing the nature of the problem that decides on the playfulness and engagement level of the game.

6 Conclusions and Future Work

The focus of this research has been primarily on investigating gamification and games with a purpose for facilitating data gathering processes for named entity disambiguation (NED). Despite the fact that gamification still is on its infancy stage, numerous benefits and advantages can be accounted by applying this concept into complex systems. When applying gamification, the main challenge is how to create a playful environment out of a complex and mundane system regardless of its nature. Appropriate and useful game design must keep the player entertained. In addition to that, it must shift player's focus towards the main and original task which is a reliable solutions to high complexity problems. This research study provides a solution to this problem through gamification. Analysis and results of the data gathered from experimental studies show that all game design decisions successfully contributed to a playful and engaging game. Additionally, the game maintained the player's focus towards providing reliable solution to the NED problem. Three research questions were used as a guide for assessing the validity and reliability of the proposed solution to gamification of NED.

What are the underlying qualities of a NED framework for supporting the generation of trustful annotation data by ordinary non-expert annotators? How motivated are users in performing annotations using the non-gamified version of the framework?

In order to test the effectiveness and usability of information generated by the framework, experimental user studies have been used to collect statistical data for analysis. Statistically significant results indicate that the annotation data generated by the non-expert annotators using the framework constitute of high quality with an accuracy (f-score) of 0.92. This corresponding performance comes as a result of a highly scalable, loosely coupled microservice framework that utilizes state-of-art components of NLP for getting the best out of NED. A framework that accurately recognizes entity mentions in text, generates useful contextual information to assist disambiguation and is easily configurable for other sorts of NLP problems without re-configuring the complete framework are the main qualities of our system. Additionally, the acquired performance results significantly outperform the best state-of-art automatic approaches to NED by 42% on the Spotlight and KORE50 datasets. Average participant agreement level of 51% on a 0.11% probability chance for correct random agreement support the effectiveness of the generated information for correct disambiguation by human annotators.

Despite the fact that the non-gamified interface used on the first evaluation experiment conforms to some design patterns, results reported seldom to moderate level of frustration. Participants perceived the interface somewhat engaging with a small number of participants rating their experience with the interface (task) as normal and tedious. Although, the non-gamified interface was perceived as interesting, somewhat engaging with a slight possibility to perform the task with the same interface in the future, results of a third questionnaire contradict such claims. 87% of participants that took part in both experiments reported to have preferred the game for performing annotation. Therefore, using only the qualities of the framework without applying gamification as a motivating mechanism, retaining users for contribution can be nearly impossible without any form of payment incentive.

How much can proximity context features such as bigrams, neighbor entities and topic keywords contribute to informing human annotators for making correct disambiguation?

The appropriate formulation of context solemnly depends on the task and target (human or machine) that uses it. This research study was concerned in finding features that would be appropriate and helpful for non-expert users to easily grasp the context in which the entity occurs. Thus, correctly disambiguate it with the appropriate candidate. Proximity features such as bi-gram collocations, neighbor entity mentions and topical document keywords have been generated by the framework as contextual information. Statistical analysis indicate that 66.57% of participants preferred these features as appropriate contextual clues for helping in disambiguation of entities compared to complete sentences. Furthermore, agreement levels reported from the previous research question also partially support the usefulness of such clues. Statistical analysis on the other hand, indicate insignificant differences between the group who preferred the short clues and the one that preferred sentences. Therefore, confident conclusions cannot be drawn based on the data and statistics acquired. Further experimental studies are necessary to confidently answer this research question.

What game mechanics can be employed in the entity disambiguation task so that high levels of engagement are achieved while still maintaining annotation quality? How do they affect player intrinsic motivation?

Different complex systems require specific game design that targets the solution of the problem being addressed. Despite the fact that the answer to this research question is primarily targeting entity disambiguation systems, the game mechanics employed here can be utilized to other systems from a conceptual point of view. Statistical significant results from the analysis of the engagement level and playfulness of the game indicate that the following game mechanics contributed to creating a successful GWAP for NED:

- Onboarding
- Triangularity (supporting both players who like to take risks and those who like to play safe)
- Non-Formal and visually appealing feedback
- Appropriately and dynamically adjusting game complexity as the player advances through the levels (maintaining game flow)
- Controlling for malicious player behaviour and maintaining gameplay quality
- Social interaction mechanisms within the game
- A well defined game engagement loop which motivates emotion, assures re-engagement with the game, provides social call to action and makes the progress of players socially visible

Results indicate an overall 95% accuracy of annotation for the entities resolved during the experimental study. Statistical significant differences for $p = 0.05$ and $p = 0.01$ in both perceived engagement and playfulness/attractiveness between the game and the plain (non-gamified) interface support the effectiveness of the employed game mechanics. Additionally, game elements such as onboarding, game instructions, the number of game options to choose from and game categories positively affected psychological needs for competence. Autonomy was positively affected as a result of having a unique and attractive game with engaging game elements and materials. Being perceived as highly enjoyable and interactive, the game accounts for additional positive affection of autonomy. Feelings of relatedness were positively affected by the social factor incorporated in the game through challenges, bets and leaderboards.

Besides the statistical analysis, the results and the answers to the posed research questions, this study constitutes of work carried out in solving architectural and design decisions composing the complete gamified system. Each design decision made on choosing specific game elements or framework algorithms is motivated based on empirical research or usability best practices. Although the primary focus of the study is revolved around gamification, additional contributions can be accounted. The implementation of the framework together with all architectural and design decisions contribute a great deal to the overall quality of this research.

The work and results reported by this research study have theoretical and practical implications in the respective field. This research study contributes to perishing the doubt of GWAP being successfully and effectively applied in non-gaming contexts. As long as the design of the game is based on empirical foundations and other psychological factors that impact human intrinsic motivation, GWAP will remain an efficient approach for harvesting the power of human computation. Providing novel approaches for effectively and efficiently improving named entity disambiguation systems results in continuous improvements in the areas of information retrieval systems and semantic web [5]. It is this sort of research work that devises techniques which can substantially improve efficiency and scalability while retaining high quality and accuracy of data.

Future Work

Regarding future work, improvements on some specific game elements and additional tests that would strengthen the results of this research have to be considered. Further improvements on the technique of measuring game complexity would result in potential improvements on the overall engagement and playfulness of the game. Calculating the word complexity within a paragraph instead of increasing the paragraph length as a measure of game complexity during fast typing would be something interesting to implement and test in future work. Additionally, the betting feature of the game was designed with the purpose of assessing annotation quality. However, the results of the study do not provide enough evidence for supporting this claim. Running the game on long time periods and assessing the effectiveness of the betting feature with regard to player annotation quality is an additional test left for future work. An interesting follow-up research would be investigating on how well would the proposed approach fit for the problem of WSD. Additionally, testing the effectiveness of the different game elements in maintain player motivation for longer periods of time represents another potential research direction.

Improving the evaluation methodology for assessing the effectiveness of contextual clues, as mention in the discussion section, is an additional part left for future work. An appropriate evaluation methodology would be a two-fold experimental study with one group being exposed by the contextual clues and the other with the complete sentence. Comparing the quality of generated annotations for both groups would result in a much more appropriate and valid assessment methodology. Finally, conducting an additional research study with a larger sample size of the target population would strengthen the claims, validity and reliability of this research work.

Bibliography

[1] Abele, P. McCrae, B. J. C. 2017. Linking open data cloud diagram 2017. http://lod-cloud.net/. Accessed: 2017-05-19.

[2] Siemens, J. C., Smith, S., Fisher, D., Thyroff, A., & Killian, G. 2015. Level up! the role of progress feedback type for encouraging intrinsic motivation and positive brand attitudes in public versus private gaming contexts. *Journal of Interactive Marketing*, 32, 1–12.

[3] Steinmetz, N., Knuth, M., & Sack, H. 2013. Statistical analyses of named entity disambiguation benchmarks. In *Proceedings of the 2013th International Conference on NLP & DBpedia-Volume 1064*, 91–102. CEUR-WS. org.

[4] Navigli, R. 2009. Word sense disambiguation: A survey. *ACM Computing Surveys (CSUR)*, 41(2), 10.

[5] Shen, W., Wang, J., & Han, J. 2015. Entity linking with a knowledge base: Issues, techniques, and solutions. *IEEE Transactions on Knowledge and Data Engineering*, 27(2), 443–460.

[6] Zheng, Z., Li, F., Huang, M., & Zhu, X. 2010. Learning to link entities with knowledge base. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 483–491. Association for Computational Linguistics.

[7] Mihalcea, R. & Csomai, A. 2007. Wikify!: linking documents to encyclopedic knowledge. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, 233–242. ACM.

[8] Shen, W., Wang, J., Luo, P., & Wang, M. 2012. Linden: linking named entities with knowledge base via semantic knowledge. In *Proceedings of the 21st international conference on World Wide Web*, 449–458. ACM.

[9] Guha, R., McCool, R., & Miller, E. 2003. Semantic search. In *Proceedings of the 12th international conference on World Wide Web*, 700–709. ACM.

[10] Mendes, P. N., Jakob, M., García-Silva, A., & Bizer, C. 2011. Dbpedia spotlight: shedding light on the web of documents. In *Proceedings of the 7th international conference on semantic systems*, 1–8. ACM.

[11] Bizer, C., Heath, T., & Berners-Lee, T. 2009. Linked data-the story so far. *Semantic services, interoperability and web applications: emerging concepts*, 205–227.

[12] Trani, S., Ceccarelli, D., Lucchese, C., Orlando, S., & Perego, R. 2014. Manual annotation of semi-structured documents for entity-linking. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, 2075–2077. ACM.

[13] Hinze, A., Heese, R., Luczak-Rösch, M., & Paschke, A. 2012. Semantic enrichment by non-experts: usability of manual annotation tools. *The Semantic Web–ISWC 2012*, 165–181.

[14] Von Ahn, L. & Dabbish, L. 2008. Designing games with a purpose. *Communications of the ACM*, 51(8), 58–67.

[15] Poesio, M., Chamberlain, J., Kruschwitz, U., Robaldo, L., & Ducceschi, L. 2013. Phrase detectives: Utilizing collective intelligence for internet-scale language resource creation. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 3(1), 3.

[16] Jurgens, D. & Navigli, R. 2014. It's all fun and games until someone annotates: Video games with a purpose for linguistic annotation. *Transactions of the Association for Computational Linguistics*, 2, 449–464.

[17] Khan, A., Tiropanis, T., & Martin, D. 2016. Exploiting semantic annotation of content with linked open data (lod) to improve searching performance in web repositories of multi-disciplinary research data. In *Information Retrieval*, 130–145. Springer.

[18] Du, F., Chen, Y., & Du, X. 2013. Linking entities in unstructured texts with rdf knowledge bases. In *Asia-Pacific Web Conference*, 240–251. Springer.

[19] Green, N., Breimyer, P., Kumar, V., & Samatova, N. F. 2010. Packplay: Mining semantic data in collaborative games. In *Proceedings of the Fourth Linguistic Annotation Workshop*, 227–234. Association for Computational Linguistics.

[20] Heese, R., Luczak-Rösch, M., Paschke, A., Oldakowski, R., & Streibel, O. 2010. One click annotation. In *SFSW*.

[21] Seaborn, K. & Fels, D. I. 2015. Gamification in theory and action: A survey. *International Journal of Human-Computer Studies*, 74, 14–31.

[22] Chamberlain, J., Fort, K., Kruschwitz, U., Lafourcade, M., & Poesio, M. 2013. Using games to create language resources: Successes and limitations of the approach. In *The People's Web Meets NLP*, 3–44. Springer.

[23] Ryan, R. M., Rigby, C. S., & Przybylski, A. 2006. The motivational pull of video games: A self-determination theory approach. *Motivation and emotion*, 30(4), 344–360.

[24] Bauer, F. & Kaltenböck, M. 2011. Linked open data: The essentials. *Edition mono/monochrom, Vienna*.

[25] Cai, J., Lee, W. S., & Teh, Y. W. 2007. Improving word sense disambiguation using topic features. In *EMNLP-CoNLL*, 1015–1023.

[26] Chan, S. W. 2013. Generating context templates for word sense disambiguation. In *Australasian Conference on Artificial Intelligence*, 466–477. Springer.

[27] Sanderson, M. 1994. Word sense disambiguation and information retrieval. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, 142–151. Springer-Verlag New York, Inc.

[28] Cucerzan, S. 2007. Large-scale named entity disambiguation based on wikipedia data.

[29] Lopes, C. T. 2009. Context features and their use in information retrieval. In *Third BCS-IRSG symposium on future directions in information access*.

[30] Bontas, E. P. 2004. Context representation and usage for the semantic web.

[31] Lamjiri, A. K., El Demerdash, O., & Kosseim, L. 2004. Simple features for statistical word sense disambiguation. In *Proceedings of Senseval-3: 3rd International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, 133–136.

[32] Pedersen, T. 2001. A decision tree of bigrams is an accurate predictor of word sense. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, 1–8. Association for Computational Linguistics.

[33] Lee, Y. K. & Ng, H. T. 2002. An empirical evaluation of knowledge sources and learning algorithms for word sense disambiguation. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, 41–48. Association for Computational Linguistics.

[34] Pedersen, T. & Kulkarni, A. 2005. Identifying similar words and contexts in natural language with senseclusters. In *Proceedings of the National Conference on Artificial Intelligence*, volume 20, 1694. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.

[35] Zhou, S., Kruengkrai, C., Okazaki, N., & Inui, K. 2014. Exploring linguistic features for named entity disambiguation. *International Journal of Computational Linguistics and Applications*, 5(2), 49.

[36] Mihalcea, R. 2002. Instance based learning with automatic feature selection applied to word sense disambiguation. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, 1–7. Association for Computational Linguistics.

[37] Nadeau, D. & Sekine, S. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1), 3–26.

[38] Finkel, J. R., Grenager, T., & Manning, C. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd annual meeting on association for computational linguistics*, 363–370. Association for Computational Linguistics.

[39] Milne, D. & Witten, I. H. 2008. Learning to link with wikipedia. In *Proceedings of the 17th ACM conference on Information and knowledge management*, 509–518. ACM.

[40] Ceccarelli, D., Lucchese, C., Orlando, S., Perego, R., & Trani, S. 2013. Learning relatedness measures for entity linking. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, 139–148. ACM.

[41] Kahjogh, B. O., Karimov, J., & Dogdu, E. 2016. Edl: A framework for entity disambiguation and linking to knowledgebases. In *Semantic Computing (ICSC), 2016 IEEE Tenth International Conference on*, 168–169. IEEE.

[42] Hoffart, J., Yosef, M. A., Bordino, I., Fürstenau, H., Pinkal, M., Spaniol, M., Taneva, B., Thater, S., & Weikum, G. 2011. Robust disambiguation of named entities in text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 782–792. Association for Computational Linguistics.

[43] Chabchoub, M., Gagnon, M., & Zouaq, A. 2016. Collective disambiguation and semantic annotation for entity linking and typing. In *Semantic Web Evaluation Challenge*, 33–47. Springer.

[44] van Veen, T., Lonij, J., & Faber, W. J. 2016. Linking named entities in dutch historical newspapers. In *Metadata and Semantics Research: 10th International Conference, MTSR 2016, Göttingen, Germany, November 22-25, 2016, Proceedings*, 205–210. Springer.

[45] Snow, R., O'Connor, B., Jurafsky, D., & Ng, A. Y. 2008. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the conference on empirical methods in natural language processing*, 254–263. Association for Computational Linguistics.

[46] Bontcheva, K., Derczynski, L., & Roberts, I. 2014. Crowdsourcing named entity recognition and entity linking corpora. *The Handbook of Linguistic Annotation (to appear)*.

[47] Ceccarelli, D., Lucchese, C., Orlando, S., Perego, R., & Trani, S. 2013. Dexter: an open source framework for entity linking. In *Proceedings of the sixth international workshop on Exploiting semantic annotations in information retrieval*, 17–20. ACM.

[48] Lewis, F. 2014. Microservices: a definition of this new architectural term. https://martinfowler.com/articles/microservices.html. Accessed: 2017-05-19.

[49] Kasun, I. 2016. Microservices in practice: From architecture to deployment. https://dzone.com/articles/microservices-in-practice-1. Accessed: 2017-05-19.

[50] Colson, J.-P. 2010. Automatic extraction of collocations: a new web-based method. *S. Bolasco, S., Chiari, I. & L. Giuliano, Proceedings of JADT*, 397–408.

[51] Duarte, E. F., Nanni, L. P., Geraldi, R. T., OliveiraJr, E., Feltrim, V. D., & Pereira, R. 2016. Ordering matters: An experimental study of ranking influence on results selection behavior during exploratory search.

[52] Usbeck, R., Röder, M., Ngonga Ngomo, A.-C., Baron, C., Both, A., Brümmer, M., Ceccarelli, D., Cornolti, M., Cherix, D., Eickmann, B., et al. 2015. Gerbil: general entity annotator benchmarking framework. In *Proceedings of the 24th International Conference on World Wide Web*, 1133–1143. ACM.

[53] Zichermann, G. & Cunningham, C. 2011. *Gamification by design: Implementing game mechanics in web and mobile apps*. " O'Reilly Media, Inc.".

[54] Cooper, S., Treuille, A., Barbero, J., Leaver-Fay, A., Tuite, K., Khatib, F., Snyder, A. C., Beenen, M., Salesin, D., Baker, D., et al. 2010. The challenge of

designing scientific discovery games. In *Proceedings of the Fifth international Conference on the Foundations of Digital Games*, 40–47. ACM.

[55] Sailer, M., Hense, J. U., Mayr, S. K., & Mandl, H. 2017. How gamification motivates: An experimental study of the effects of specific game design elements on psychological need satisfaction. *Computers in Human Behavior*, 69, 371–380.

[56] Mekler, E. D., Brühlmann, F., Tuch, A. N., & Opwis, K. 2015. Towards understanding the effects of individual gamification elements on intrinsic motivation and performance. *Computers in Human Behavior*.

[57] Venhuizen, N., Evang, K., Basile, V., & Bos, J. 2013. Gamification for word sense labeling. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)*.

[58] Schell, J. 2014. *The Art of Game Design: A book of lenses*. CRC Press.

[59] Ratinov, L. & Roth, D. 2009. Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, 147–155. Association for Computational Linguistics.

[60] Dong, X., Harper, F. M., & Konstan, J. A. 2011. Entity-linking interfaces in user-contributed content: preference and performance. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2187–2196. ACM.

[61] Vannella, D., Jurgens, D., Scarfini, D., Toscani, D., & Navigli, R. 2014. Validating and extending semantic knowledge bases using video games with a purpose. In *ACL (1)*, 1294–1304.

A Experiment 1 - AnnotateMe

A.1 Documentation of RabbitMQ Message Routes

The table below represents all the communication infrastructure for the asynchronous publish-subscribe messages protocol. All microservices composing the NED Framework exchange information with each other by publishing and subscribing to each-others' message queues using RabbitMQ as the underlying technology provider. All the parameters associated with each communication route (Routing Key - Column 1), are mandatory and shall be provided as a complete JSON object with the specified name and the corresponding data type.

RabbitMQ API Documentation

Routing Key	Publisher	Subscribers	Parameters	
create.document.data	Admin Front-End App	Datastore-Service	path	String
			content	String
			dataset	String
			confidence	Number
			support	Number
			nrKeywordsToExtract	Number
			nrConceptsToExtract	Number
			sentenceRestricted	Boolean
			routeKey	String
extract.document.entities&keywords	Datastore-Service	NER-Service	documentID	Number
			textData	String
			confidence	Numer
		TopicKeyword-Service	support	Number
			nrKeywordsToExtract	Number
			nrConceptsToExtract	Number
create.document.entities	NER-Service	Datastore-Service	status	Number
			text	String
			documentID	Number
			entities	JSON
			stanford	JSON
			dbpedia	JSON
			sections	JSON
			tokensWithPunctuations	JSON
			tokensByWords	JSON
			confidence	Number
			support	Number
			nrKeywordsToExtract	Number
			nrConceptsToExtract	Number
			routeKey	String
create.document.keywords	TopicKeyword-Service	Datastore-Service	documentID	Number
			keywords	JSON
			routeKey	String
extract.entity.clues	Datastore-Service	ContextClue-Service	documentID	Number
			sentences	JSON
			sentenceRestricted	Boolean
extract.entity.candidates	Datastore-Service	CandidateGenration-Service	documentID	Number
			entities	JSON
			confidence	Number
			support	Number
create.entity.clues	ContextClue-Service	Datastore-Service	routeKey	String
			entities	JSON

create.entity.candidates	CandidateGeneration-Service	Datastore-Service	documentID	Number
			entities	JSON
			routeKey	String
document.datageneration.done	Datastore-Service	Admin Front-End App	status	Number
			msg	String
			documentID	Number

A.2 Documentation of Datastore and Dataprep Microservice REST routes

This appendix documents all the routes used to access the resources residing in the **Datastore Microservice** as well as the documented routes for accessing resources from the **Dataprep Microservice**. Swagger Documentation API[1] has been used for generating the documentation schema illustrated below.

[1] Swagger IO http://swagger.io/

Swagger Datastore-Service 1.0.0

`[ Base url: datastore-service.swagger.io/v1]`

REST API route documentation for DataStore Microservice.

Schemes

HTTP

docs

GET `/docs` Gets all documents

POST `/docs` Creates a new document

PUT `/docs` Updates an existing document

GET `/docs/{docId}` Find document by id

DELETE `/docs/{docId}` Deletes a document

GET `/docs/{docId}/keywords` Finds all document keywords

POST `/docs/{docId}/keywords` Creates a new document keyword

DELETE `/docs/{docId}/keywords` Deletes all document keywords

GET `/docs/{docId}/sentances` Finds all document sentences

POST `/docs/{docId}/sentances` Creates a new document sentence

PUT `/docs/{docId}/sentances` Updates an existing document sentence

DELETE `/docs/{docId}/sentances` Deletes all document Sentences

GET `/docs/{docId}/sentances/{sentanceId}` Finds a specific document sentence

GET `/docs/{docId}/entities` Finds document entities

participants

GET `/participants` Finds all existing participants

POST `/participants` Creates a new participant

DELETE `/participants` Deletes an exisiting participant

GET `/participants/{participantId}` Finds a participants by Id

GET `/participants/{participantId}/annotations` Finds all annotations generated by participant

POST `/participants/{participantId}/updateStartTime` Updates the annotation start time of participants

POST `/participants/{participantId}/updateEndTime` Updates the annotation end time of participants

annotations

GET `/annotations` Finds all annotations

POST `/annotations` Creates a new annotation

GET `/annotations/{annotationId}` Finds annotation by id

PUT `/annotations/{annotationId}` Updates an existing annotation

	`/annotations/{annotationId}`	
DELETE	`/annotations/{annotationId}`	Deletes an existing annotation

entities

GET	`/entities`	Finds all entities
POST	`/entities`	Creates a new entity
GET	`/entities/{entityId}`	Finds an entity by id
PUT	`/entities/{entityId}`	Updates an existing entity
DELETE	`/entities/{entityId}`	Deletes an existing entitiy
PUT	`/entities/{entityId}/resolve`	Resolves/Disambiguates an existing entity
PUT	`/entities/{entityId}/threshold`	Updates the ambiguity threshold of an existing entity
GET	`/entities/{entityId}/collocations`	Finds all collocations associated with the entity
POST	`/entities/{entityId}/collocations`	Associates a new collocation with the entity
DELETE	`/entities/{entityId}/collocations`	Deletes all collocations associated with the entity
GET	`/entities/{entityId}/candidates`	Finds all candidates associated with the entity
POST	`/entities/{entityId}/candidates`	Associates a new candidate with the entity
DELETE	`/entities/{entityId}/candidates`	Deletes all candidates associated with the entity

brooker

POST `/brookerInvoke` Invokes the (amqp) message brooker that initiates the framework named entity linking procedure

game

POST `/game/authenticate` Authenticates player into the game

POST `/game/register` Registers players' authentication information into the system

GET `/game/categories/{categoryId}` Finds all game categories

POST `/game/categories/{categoryId}` Creates a new game category

GET `/game/playerStats/{playerId}` Finds game statistics of a player

POST `/game/score/{playerId}` Persists new scores to the payer

POST `/game/wpm/{playerId}` Persists new wpm score for the player

POST `/game/level/{playerId}/levelUpPlayer` Checks (based on accumulated points) if player has leveled up and assigns a new level to the player

POST `/game/addGameRound` Creates a new game round (annotation version for the game)

GET `/game/getChallengers/{wpm}/player/{playerId}` Finds possible challengers for player to challenge based on WPM

POST `/game/challengePlayers` Creates a new game challenge

GET `/game/getPlayerChallenges/{playerId}` Finds all (pending) challanges assigned to the player

GET `/game/getChallengeInfo/{challengeId}` Fetches game challenge information

Updates a game challenge, deciding who the looser and the winner of

POST `/game/updateChallenge` the challenge

GET `/game/getProfileStats/{playerId}` Finds players' profile information

GET `/game/getUpdatedChallenges/{playerId}` Fetches all challenges that have been completed (Won, Lost, Draw)

GET `/game/getLeaderBoard` Fetches the leaderbord with all players and theircorresponding rankings

GET `/game/getPlayerPositionInLeaderboard/{playerId}` Finds the players' exact position on the leaderboard

Swagger Dataprep-Service 1.0.0

`[ Base url: dataprep-service.swagger.io/v1]`

REST API route documentation for Data-Preparation Microservice.

Schemes

HTTP

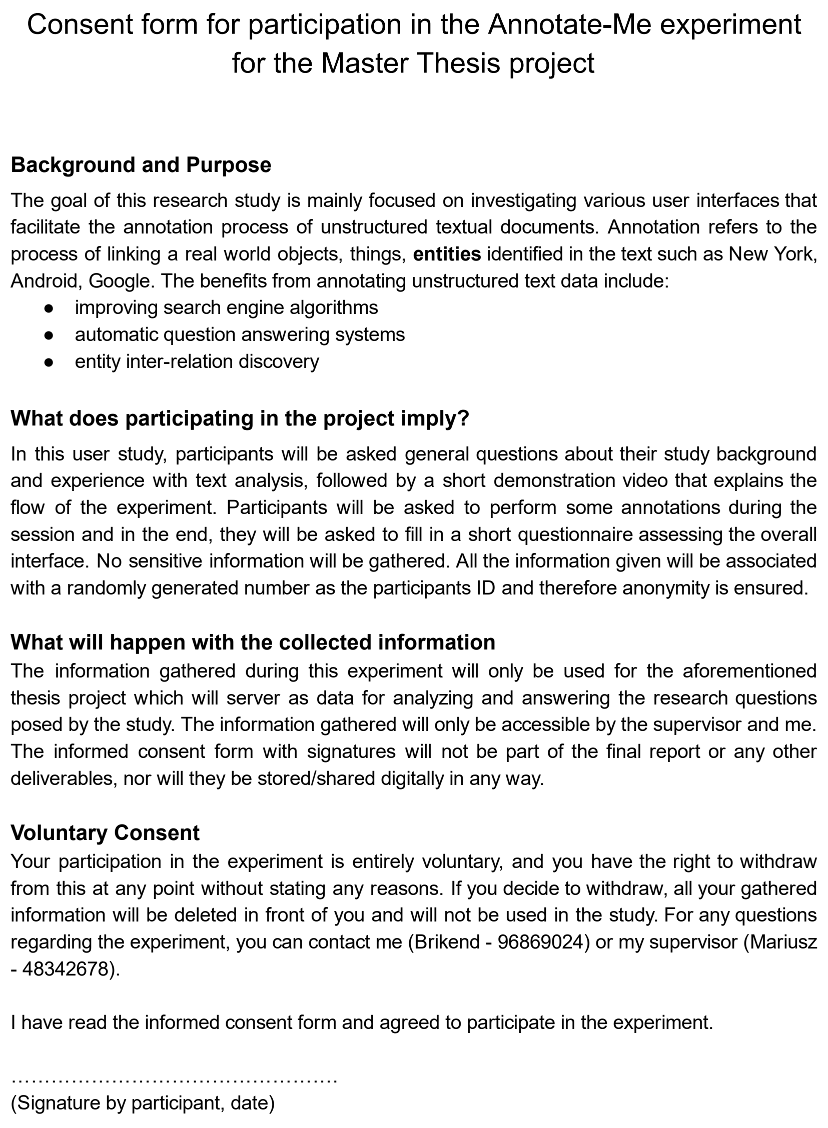
annotation-api

GET `/annotateme/api` Prepares all necessary data required by the AnnotateMe Interface

game-api

GET `/game/api` Prepares all necessary data required by the Fastype Game

A.3 Consent Form



Consent form for participation in the Annotate-Me experiment for the Master Thesis project

Background and Purpose

The goal of this research study is mainly focused on investigating various user interfaces that facilitate the annotation process of unstructured textual documents. Annotation refers to the process of linking a real world objects, things, **entities** identified in the text such as New York, Android, Google. The benefits from annotating unstructured text data include:

- improving search engine algorithms
- automatic question answering systems
- entity inter-relation discovery

What does participating in the project imply?

In this user study, participants will be asked general questions about their study background and experience with text analysis, followed by a short demonstration video that explains the flow of the experiment. Participants will be asked to perform some annotations during the session and in the end, they will be asked to fill in a short questionnaire assessing the overall interface. No sensitive information will be gathered. All the information given will be associated with a randomly generated number as the participants ID and therefore anonymity is ensured.

What will happen with the collected information

The information gathered during this experiment will only be used for the aforementioned thesis project which will server as data for analyzing and answering the research questions posed by the study. The information gathered will only be accessible by the supervisor and me. The informed consent form with signatures will not be part of the final report or any other deliverables, nor will they be stored/shared digitally in any way.

Voluntary Consent

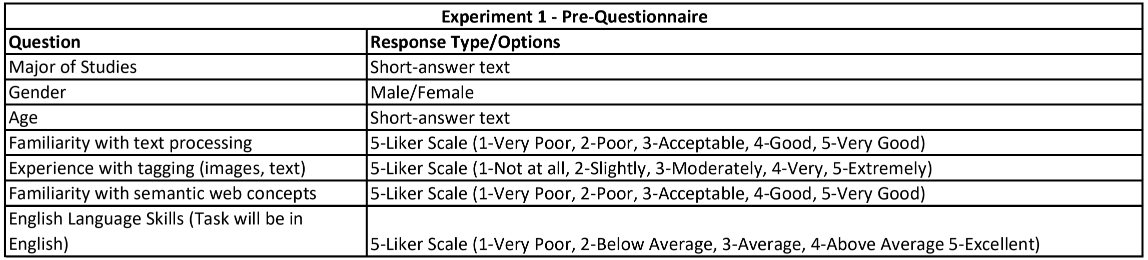
Your participation in the experiment is entirely voluntary, and you have the right to withdraw from this at any point without stating any reasons. If you decide to withdraw, all your gathered information will be deleted in front of you and will not be used in the study. For any questions regarding the experiment, you can contact me (Brikend - 96869024) or my supervisor (Mariusz - 48342678).

I have read the informed consent form and agreed to participate in the experiment.

…………………………………………

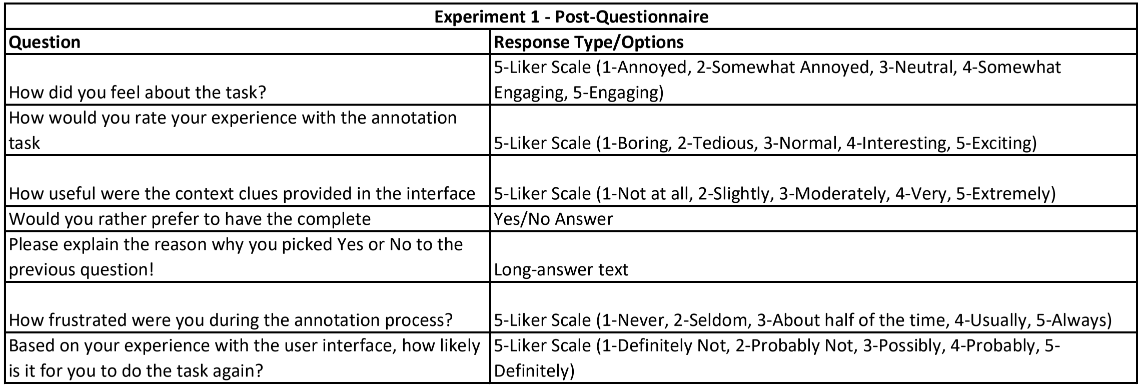
(Signature by participant, date)

A.4 Pre-Questionnaire



Experiment 1 - Pre-Questionnaire	
Question	**Response Type/Options**
Major of Studies	Short-answer text
Gender	Male/Female
Age	Short-answer text
Familiarity with text processing	5-Liker Scale (1-Very Poor, 2-Poor, 3-Acceptable, 4-Good, 5-Very Good)
Experience with tagging (images, text)	5-Liker Scale (1-Not at all, 2-Slightly, 3-Moderately, 4-Very, 5-Extremely)
Familiarity with semantic web concepts	5-Liker Scale (1-Very Poor, 2-Poor, 3-Acceptable, 4-Good, 5-Very Good)
English Language Skills (Task will be in English)	5-Liker Scale (1-Very Poor, 2-Below Average, 3-Average, 4-Above Average 5-Excellent)

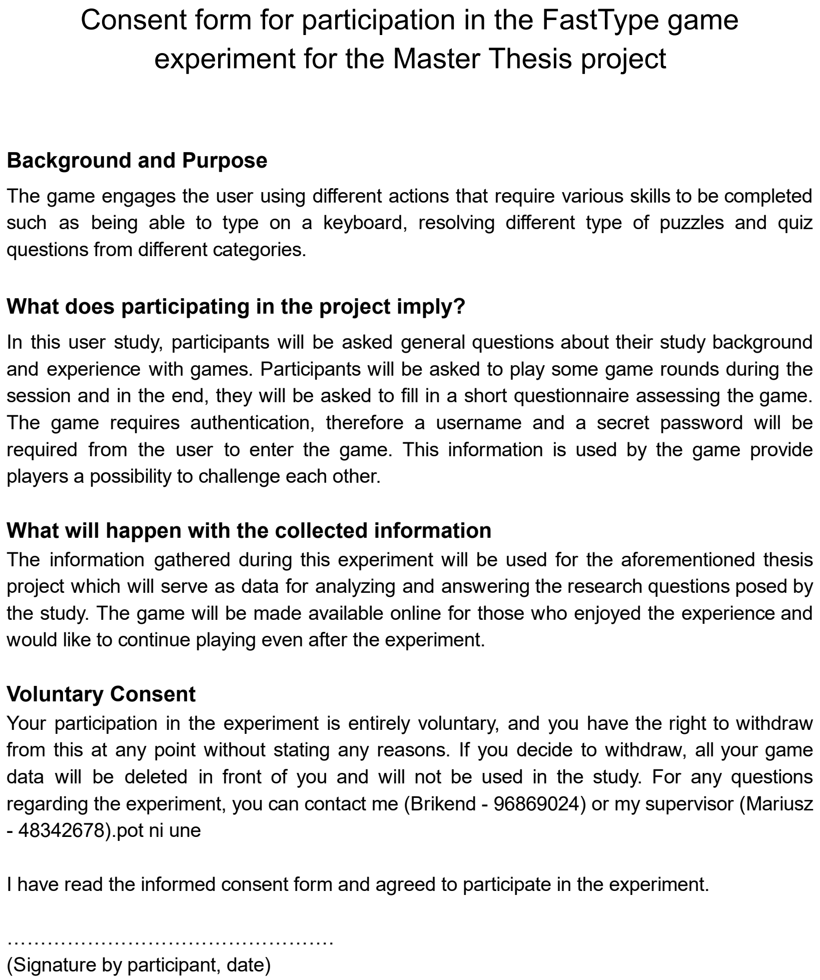
A.5 Post-Questionnaire



Experiment 1 - Post-Questionnaire	
Question	**Response Type/Options**
How did you feel about the task?	5-Liker Scale (1-Annoyed, 2-Somewhat Annoyed, 3-Neutral, 4-Somewhat Engaging, 5-Engaging)
How would you rate your experience with the annotation task	5-Liker Scale (1-Boring, 2-Tedious, 3-Normal, 4-Interesting, 5-Exciting)
How useful were the context clues provided in the interface	5-Liker Scale (1-Not at all, 2-Slightly, 3-Moderately, 4-Very, 5-Extremely)
Would you rather prefer to have the complete	Yes/No Answer
Please explain the reason why you picked Yes or No to the previous question!	Long-answer text
How frustrated were you during the annotation process?	5-Liker Scale (1-Never, 2-Seldom, 3-About half of the time, 4-Usually, 5-Always)
Based on your experience with the user interface, how likely is it for you to do the task again?	5-Liker Scale (1-Definitely Not, 2-Probably Not, 3-Possibly, 4-Probably, 5-Definitely)

B Experiment 2 - Fastype

B.1 Consent Form



Consent form for participation in the FastType game experiment for the Master Thesis project

Background and Purpose

The game engages the user using different actions that require various skills to be completed such as being able to type on a keyboard, resolving different type of puzzles and quiz questions from different categories.

What does participating in the project imply?

In this user study, participants will be asked general questions about their study background and experience with games. Participants will be asked to play some game rounds during the session and in the end, they will be asked to fill in a short questionnaire assessing the game. The game requires authentication, therefore a username and a secret password will be required from the user to enter the game. This information is used by the game provide players a possibility to challenge each other.

What will happen with the collected information

The information gathered during this experiment will be used for the aforementioned thesis project which will serve as data for analyzing and answering the research questions posed by the study. The game will be made available online for those who enjoyed the experience and would like to continue playing even after the experiment.

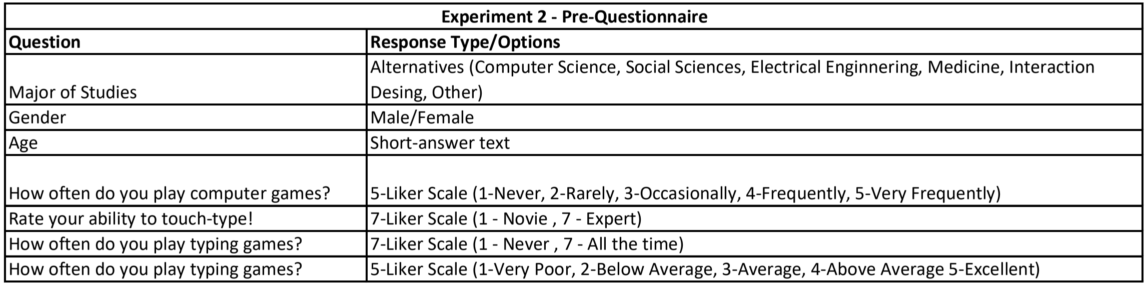
Voluntary Consent

Your participation in the experiment is entirely voluntary, and you have the right to withdraw from this at any point without stating any reasons. If you decide to withdraw, all your game data will be deleted in front of you and will not be used in the study. For any questions regarding the experiment, you can contact me (Brikend - 96869024) or my supervisor (Mariusz - 48342678).pot ni une

I have read the informed consent form and agreed to participate in the experiment.

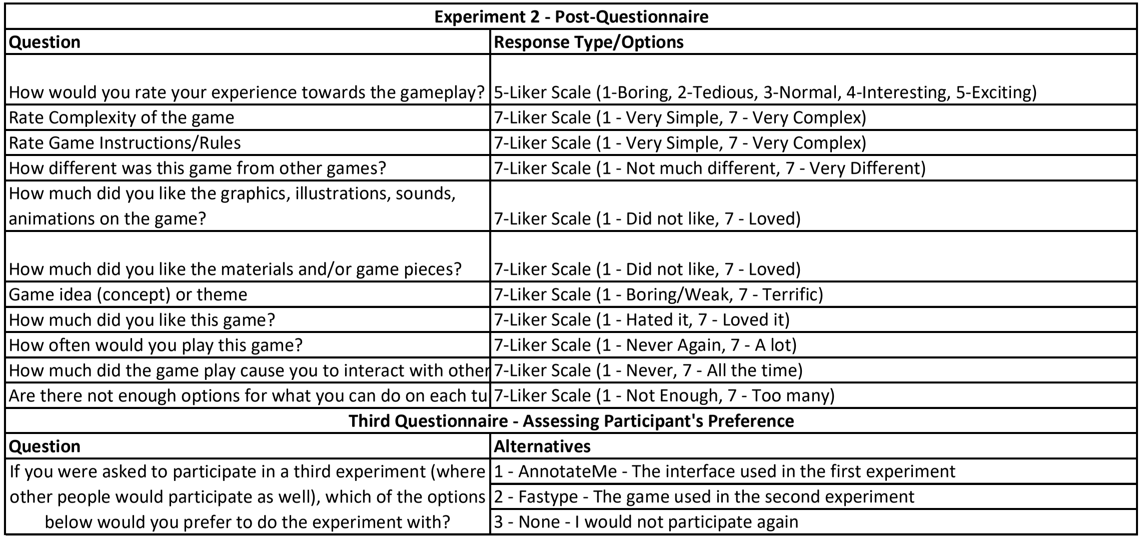
…………………………………………
(Signature by participant, date)

B.2 Pre-Questionnaire



Experiment 2 - Pre-Questionnaire	
Question	**Response Type/Options**
Major of Studies	Alternatives (Computer Science, Social Sciences, Electrical Enginnering, Medicine, Interaction Desing, Other)
Gender	Male/Female
Age	Short-answer text
How often do you play computer games?	5-Liker Scale (1-Never, 2-Rarely, 3-Occasionally, 4-Frequently, 5-Very Frequently)
Rate your ability to touch-type!	7-Liker Scale (1 - Novie , 7 - Expert)
How often do you play typing games?	7-Liker Scale (1 - Never , 7 - All the time)
How often do you play typing games?	5-Liker Scale (1-Very Poor, 2-Below Average, 3-Average, 4-Above Average 5-Excellent)

B.3 Post-Questionnaire & Preference Questionnaire



Experiment 2 - Post-Questionnaire	
Question	**Response Type/Options**
How would you rate your experience towards the gameplay?	5-Liker Scale (1-Boring, 2-Tedious, 3-Normal, 4-Interesting, 5-Exciting)
Rate Complexity of the game	7-Liker Scale (1 - Very Simple, 7 - Very Complex)
Rate Game Instructions/Rules	7-Liker Scale (1 - Very Simple, 7 - Very Complex)
How different was this game from other games?	7-Liker Scale (1 - Not much different, 7 - Very Different)
How much did you like the graphics, illustrations, sounds, animations on the game?	7-Liker Scale (1 - Did not like, 7 - Loved)
How much did you like the materials and/or game pieces?	7-Liker Scale (1 - Did not like, 7 - Loved)
Game idea (concept) or theme	7-Liker Scale (1 - Boring/Weak, 7 - Terrific)
How much did you like this game?	7-Liker Scale (1 - Hated it, 7 - Loved it)
How often would you play this game?	7-Liker Scale (1 - Never Again, 7 - A lot)
How much did the game play cause you to interact with other	7-Liker Scale (1 - Never, 7 - All the time)
Are there not enough options for what you can do on each tu	7-Liker Scale (1 - Not Enough, 7 - Too many)
Third Questionnaire - Assessing Participant's Preference	
Question	**Alternatives**
If you were asked to participate in a third experiment (where other people would participate as well), which of the options below would you prefer to do the experiment with?	1 - AnnotateMe - The interface used in the first experiment
	2 - Fastype - The game used in the second experiment
	3 - None - I would not participate again

C ANOVA - Raw Data Analysis

C.1 Attractiveness of Interfaces

Below we present the ANOVA calculations performed on the gathered observations when asked participants about their experience with both interfaces.

	ANOVA - Observations of Participants Experience Towards Both Interfaces (Fastype and AnnotateMe)					
	Experiment 2 Observations (Game)	x-mean	(x-mean)^2	Experiment 1 Observations (AnnotateMe)	x-mean	(x-mean)^2
	5	0.3703703704	0.1371742112	4	0.03225806452	0.001040582726
	5	0.3703703704	0.1371742112	4	0.03225806452	0.001040582726
	4	-0.6296296296	0.3964334705	4	0.03225806452	0.001040582726
	4	-0.6296296296	0.3964334705	4	0.03225806452	0.001040582726
	4	-0.6296296296	0.3964334705	5	1.032258065	1.065556712
	5	0.3703703704	0.1371742112	5	1.032258065	1.065556712
	5	0.3703703704	0.1371742112	4	0.03225806452	0.001040582726
	4	-0.6296296296	0.3964334705	3	-0.9677419355	0.9365244537
	5	0.3703703704	0.1371742112	3	-0.9677419355	0.9365244537
	5	0.3703703704	0.1371742112	4	0.03225806452	0.001040582726
	5	0.3703703704	0.1371742112	4	0.03225806452	0.001040582726
	5	0.3703703704	0.1371742112	4	0.03225806452	0.001040582726
	5	0.3703703704	0.1371742112	3	-0.9677419355	0.9365244537
	4	-0.6296296296	0.3964334705	4	0.03225806452	0.001040582726
	4	-0.6296296296	0.3964334705	4	0.03225806452	0.001040582726
	4	-0.6296296296	0.3964334705	4	0.03225806452	0.001040582726
	4	-0.6296296296	0.3964334705	4	0.03225806452	0.001040582726
	5	0.3703703704	0.1371742112	4	0.03225806452	0.001040582726
	5	0.3703703704	0.1371742112	4	0.03225806452	0.001040582726
	5	0.3703703704	0.1371742112	4	0.03225806452	0.001040582726
	5	0.3703703704	0.1371742112	4	0.03225806452	0.001040582726
	5	0.3703703704	0.1371742112	4	0.03225806452	0.001040582726
	5	0.3703703704	0.1371742112	5	1.032258065	1.065556712
	5	0.3703703704	0.1371742112	5	1.032258065	1.065556712
	5	0.3703703704	0.1371742112	4	0.03225806452	0.001040582726
	4	-0.6296296296	0.3964334705	4	0.03225806452	0.001040582726
	4	-0.6296296296	0.3964334705	4	0.03225806452	0.001040582726
				4	0.03225806452	0.001040582726
				2	-1.967741935	3.872008325
				4	0.03225806452	0.001040582726
				4	0.03225806452	0.001040582726
Sum	125	0	6.296296296	123	0	10.96774194
Mean	4.62962963	0	0.2331961591	3.967741935	0	0.353798127
Sum of squares within groups	17.26403823					
	Observations					
	5	0.724137931	0.5243757432		**Total Sum of Squares**	**23.5862069**
	5	0.724137931	0.5243757432		**Sum of Squares Within**	**17.26403823**
	4	-0.275862069	0.07609988109		**Som uf Squares Between**	**6.322168665**
	4	-0.275862069	0.07609988109		**Degrees of Freedom**	
	4	-0.275862069	0.07609988109			
	5	0.724137931	0.5243757432		**Numerator**	**1**
	5	0.724137931	0.5243757432		**Denominator**	**56**
	4	-0.275862069	0.07609988109		**SSB/DF1**	**6.322168665**
	5	0.724137931	0.5243757432		**SSW/DF2**	**0.308286397**
	5	0.724137931	0.5243757432		**F-Score, F(1,56)**	**20.50745257**
	5	0.724137931	0.5243757432		**Critical Value, p < 0.05**	**4.002**
	5	0.724137931	0.5243757432		**Result**	**Statistically Significant for p = 0.05**
	5	0.724137931	0.5243757432		**Critical Value, p < 0.01**	7.007
	4	-0.275862069	0.07609988109		**Result**	**Statistically Significant for p = 0.01**
	4	-0.275862069	0.07609988109			
	4	-0.275862069	0.07609988109			
	4	-0.275862069	0.07609988109			
	5	0.724137931	0.5243757432			
	5	0.724137931	0.5243757432			
	5	0.724137931	0.5243757432			
	5	0.724137931	0.5243757432			
	5	0.724137931	0.5243757432			
	5	0.724137931	0.5243757432			
	5	0.724137931	0.5243757432			
	5	0.724137931	0.5243757432			
	4	-0.275862069	0.07609988109			
	4	-0.275862069	0.07609988109			
	4	-0.275862069	0.07609988109			
	4	-0.275862069	0.07609988109			
	4	-0.275862069	0.07609988109			
	4	-0.275862069	0.07609988109			
	5	0.724137931	0.5243757432			

	5	0.724137931	0.5243757432			
	4	-0.275862069	0.07609988109			
	3	-1.275862069	1.627824019			
	3	-1.275862069	1.627824019			
	4	-0.275862069	0.07609988109			
	4	-0.275862069	0.07609988109			
	4	-0.275862069	0.07609988109			
	3	-1.275862069	1.627824019			
	4	-0.275862069	0.07609988109			
	4	-0.275862069	0.07609988109			
	4	-0.275862069	0.07609988109			
	4	-0.275862069	0.07609988109			
	4	-0.275862069	0.07609988109			
	4	-0.275862069	0.07609988109			
	4	-0.275862069	0.07609988109			
	4	-0.275862069	0.07609988109			
	4	-0.275862069	0.07609988109			
	5	0.724137931	0.5243757432			
	5	0.724137931	0.5243757432			
	4	-0.275862069	0.07609988109			
	4	-0.275862069	0.07609988109			
	4	-0.275862069	0.07609988109			
	4	-0.275862069	0.07609988109			
	2	-2.275862069	5.179548157			
	4	-0.275862069	0.07609988109			
	4	-0.275862069	0.07609988109			
Mean	4.275862069					
		Total Sum of Squ	23.5862069			

C.2 Player Engagement Analysis

The analysis presented below represent the ANOVA analysis performed on the gathered observations when participants where asked to rate their engagement with both interfaces using a liker-scale measure[1]. Please note that different measure matrices were used in each experiment, however, a normalization procedure is performed for the game observation to scale the values down from a 7-liker scale to 5.

[1] For the first experiment we used 5-liker scale whereas for the second experiment we used a 7-liker scale. We normalize the observations afterwards to accurately perform comparison measures

	ANOVA - Observations of Participants Percieved Engagement with both interfaces (Fastype and AnnotateMe)						
	Experiment 2 Observations (Game)	**Normalized to Max 5**	**x-mean**	**(x-mean)^2**	**Experiment 1 Observations (AnnotateMe)**	**x-mean**	**(x-mean)^2**
	5	3.571428571	0.05291005291	0.002799473699	4	0.1481481481	0.0219478738
	5	3.571428571	0.05291005291	0.002799473699	3	-0.8518518519	0.7256515775
	4	2.857142857	-0.6613756614	0.4374177655	4	0.1481481481	0.0219478738
	4	2.857142857	-0.6613756614	0.4374177655	3	-0.8518518519	0.7256515775
	6	4.285714286	0.7671957672	0.5885893452	4	0.1481481481	0.0219478738
	6	4.285714286	0.7671957672	0.5885893452	5	1.148148148	1.31824417
	5	3.571428571	0.05291005291	0.002799473699	5	1.148148148	1.31824417
	5	3.571428571	0.05291005291	0.002799473699	3	-0.8518518519	0.7256515775
	7	5	1.481481481	2.19478738	3	-0.8518518519	0.7256515775
	5	3.571428571	0.05291005291	0.002799473699	3	-0.8518518519	0.7256515775
	7	5	1.481481481	2.19478738	3	-0.8518518519	0.7256515775
	4	2.857142857	-0.6613756614	0.4374177655	5	1.148148148	1.31824417
	5	3.571428571	0.05291005291	0.002799473699	4	0.1481481481	0.0219478738
	3	2.142857143	-1.375661376	1.89244422	4	0.1481481481	0.0219478738
	1	0.7142857143	-2.804232804	7.86372162	4	0.1481481481	0.0219478738
	6	4.285714286	0.7671957672	0.5885893452	4	0.1481481481	0.0219478738
	5	3.571428571	0.05291005291	0.002799473699	4	0.1481481481	0.0219478738
	5	3.571428571	0.05291005291	0.002799473699	4	0.1481481481	0.0219478738
	4	2.857142857	-0.6613756614	0.4374177655	3	-0.8518518519	0.7256515775
	6	4.285714286	0.7671957672	0.5885893452	4	0.1481481481	0.0219478738
	4	2.857142857	-0.6613756614	0.4374177655	5	1.148148148	1.31824417
	7	5	1.481481481	2.19478738	4	0.1481481481	0.0219478738
	7	5	1.481481481	2.19478738	5	1.148148148	1.31824417
	6	4.285714286	0.7671957672	0.5885893452	3	-0.8518518519	0.7256515775
	4	2.857142857	-0.6613756614	0.4374177655	3	-0.8518518519	0.7256515775
	1	0.7142857143	-2.804232804	7.86372162	5	1.148148148	1.31824417
	6	4.285714286	0.7671957672	0.5885893452	3	-0.8518518519	0.7256515775
					2	-1.851851852	3.429355281
					4	0.1481481481	0.0219478738
					4	0.1481481481	0.0219478738
					4	0.1481481481	0.0219478738
Sum		95	0	32.57747543	118	-1.407407407	18.90260631
Mean		3.518518519	0	1.206573164	3.851851852	0	0.5706447188
Sum of Squares Within	51.48008174						
		Observations	x-mean	(x-mean)^2			
		5	0.6724137931	0.4521403092		**Total Sum of Squares**	**100.7758621**
		5	0.6724137931	0.4521403092		**Sum of Squares Within**	**51.48008174**
		4	-0.3275862069	0.1073127229		**Sum of Squares Between**	**49.29578032**
		4	-0.3275862069	0.1073127229		**Degrees of Freedom**	
		6	1.672413793	2.796967895			
		6	1.672413793	2.796967895		**Numerator**	**1**
		5	0.6724137931	0.4521403092		**Denominator**	**56**
		5	0.6724137931	0.4521403092		**SSB/DF1**	**49.29578032**
		7	2.672413793	7.141795482		**SSW/DF2**	**0.919287174**
		5	0.6724137931	0.4521403092		**F-Score, F(1,56)**	**53.62391831**
		7	2.672413793	7.141795482		**Critical Value, p < .05**	**4.002**
		4	-0.3275862069	0.1073127229		**Result**	**Statistically Significant for p = 0.05**
		5	0.6724137931	0.4521403092		Critical Value, p < .01	7.007
		3	-1.327586207	1.762485137		**Result**	**Statistically Significant for p = 0.01**
		1	-3.327586207	11.07282996			
		6	1.672413793	2.796967895			
		5	0.6724137931	0.4521403092			
		5	0.6724137931	0.4521403092			
		4	-0.3275862069	0.1073127229			
		6	1.672413793	2.796967895			
		4	-0.3275862069	0.1073127229			
		7	2.672413793	7.141795482			
		7	2.672413793	7.141795482			
		6	1.672413793	2.796967895			
		4	-0.3275862069	0.1073127229			
		1	-3.327586207	11.07282996			
		6	1.672413793	2.796967895			
		4	-0.3275862069	0.1073127229			
		3	-1.327586207	1.762485137			
		4	-0.3275862069	0.1073127229			
		3	-1.327586207	1.762485137			
		4	-0.3275862069	0.1073127229			
		5	0.6724137931	0.4521403092			
		5	0.6724137931	0.4521403092			
		3	-1.327586207	1.762485137			
		3	-1.327586207	1.762485137			
		3	-1.327586207	1.762485137			
		3	-1.327586207	1.762485137			
		5	0.6724137931	0.4521403092			
		4	-0.3275862069	0.1073127229			
		4	-0.3275862069	0.1073127229			

		4	-0.3275862069	0.1073127229			
		4	-0.3275862069	0.1073127229			
		4	-0.3275862069	0.1073127229			
		4	-0.3275862069	0.1073127229			
		3	-1.327586207	1.762485137			
		4	-0.3275862069	0.1073127229			
		5	0.6724137931	0.4521403092			
		4	-0.3275862069	0.1073127229			
		5	0.6724137931	0.4521403092			
		3	-1.327586207	1.762485137			
		3	-1.327586207	1.762485137			
		5	0.6724137931	0.4521403092			
		3	-1.327586207	1.762485137			
		2	-2.327586207	5.417657551			
		4	-0.3275862069	0.1073127229			
		4	-0.3275862069	0.1073127229			
		4	-0.3275862069	0.1073127229			
	Mean	4.327586207					
			Total Sum of Squ	100.7758621			