Master Thesis

Entity Linking for Company Name Disambiguation

Jona B. Bosman

a thesis submitted in partial fulfilment of the requirements for the degree of

MA Linguistics
(Text Mining)

Vrije Universiteit Amsterdam

Computational Lexicology and Terminology Lab
Department of Language and Communication
Faculty of Humanities

Supervised by:	Hennie van Vliet, Lieke Gelderloos
2^{nd} reader:	Lisa Beinborn
Submitted:	June 29, 2020

Abstract

Abstract This thesis presents an Entity Linking system to link mentions of companies in news articles to unique identifiers in a Knowledge Base. The project was executed during an internship at software company Brainial. The system's design makes use of the Entity Linking component from spaCy and is characterized as a pipeline architecture, handling the three different stages of Entity Linking (Mention Detection, Candidate Generation and Entity Disambiguation) in successive order. The spaCy model is a feed-forward network trained to link the context of a mention to an entity description. As part of the project, an annotated dataset was created, on which the system was trained and evaluated. It received an F-score of 0.946 on an evaluation set. The created system greatly outperformed two baselines based on Brainial's current Entity Linking approach, but achieved comparable results with a majority baseline. This suggests that the task of Company Name Disambiguation might not be as hard as expected and that mentions of companies in news articles usually refer to the same entity, regardless of the context they appear in.

Keywords Entity Linking, Named Entity Linking, Entity Disambiguation, Company Name Disambiguation

Declaration of Authorship

I, Jona Benja Bosman, declare that this thesis, titled *Entity Linking for Company Name Disambiguation* and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a Master degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Acknowledgments

Whilst completing this thesis I received a great deal of help from many people.

First I want to thank my two supervisors, Lieke Gelderloos and Hennie van der Vliet. I am grateful that you could answer my questions, for steering me in the right direction and the wonderful feedback you gave me.

Another major contribution was made by Bob Maks, the supervisor of my internship at Brainial. Thank you for the long meetings and brainstorm sessions, and for helping me make plannings and stick to them.

My thanks goes out to Brainial for believing in me and offering me an internship on such short notice, and to Fedor, Taco, Quinten, Merlijn, Rutger, Teodor and Bob for welcoming me in the team so generously. I had a fantastic time. Also thanks to Myrthe for bringing me in contact with Brainial and to Rogier and Pepijn for contributing to the annotation process.

I am grateful for Isa Maks, for organizing the whole internship-thesis process and thank you Lisa Beinborn for being my second reader.

Jona Bosman

List of Tables

Contents

Chapter 1

Introduction

Names of companies can be ambiguous: some companies are called the same and other companies go by variations of their names, such as abbreviations. For some tasks that involve retrieving information about a company that is mentioned, it is necessary to know exactly which company is mentioned. Using just the name might not be enough to disambiguate the company, as the context of company mention (the rest of the text in which the company name appears) generally plays an important role. Entity Linking can make it possible to disambiguate between company mentions, making use of their local context. This thesis proposes an Entity Linking system for the task of company name disambiguation.

1.1 Introduction to Entity Linking

Entity Linking (EL) is the task of linking mentions of Named Entities in texts to entities in a reference set, as defined by Han et al. (2005), and is also known as Named Entity Linking (NEL), Named Entity Disambiguation (NED) and Named Entity Normalization (NEN). Named Entities refer to objects in the real world, such as persons, locations or companies and generally also temporal (e.g. dates) and numerical values (e.g. money) are included (Nadeau & Sekine, 2007). There are three main difficulties to Named Entity Linking, according to (Rao et al., 2013):

– **Name Variations**. Named entities in a text (or *mentions*) that refer to the same entity can appear under different forms (such as *Ludwig van Beethoven* or simply *Beethoven*).

– **Entity Ambiguity**. Some *mentions* have the same form but refer to a different entity (*Paris*, capital of France, and *Paris*, a city in Arkansas). The entities in the reference set are often listed as unique identifiers (e.g. numbers) to eliminate any confusion rooted in name similarity (as highlighted in the *Paris* example).

– **Absence**. It may be the case that the entity a mention refers to is not present in the Knowledge Base, thus it is not possible to link the mention, and the mention should be labelled NIL (not-in-lexicon). Systems need to be specifically trained to predict the NIL-value. In this paper, predicting the NIL-value will be out of consideration.

Systems designed to overcome these challenges can improve information retrieval (Hasibi et al., 2016). The reference entity set is sometimes referred to as a *Knowledge Base*

(Shen et al., 2014) or *Knowledge Graph* (Mulang' et al., 2020). A popular task for EL is to link entity mentions in texts, such as Tweets or blogposts, to their corresponding Wikipedia page (Hachey et al., 2013; Qureshi et al., 2014; Tsai & Roth, 2016). The task has its own name: *wikification* (Mihalcea & Csomai, 2007). In the task, the information on Wikipedia pages is used as a Knowledge Base.

1.2 Company Name Disambiguation

Companies are often referred to with variations of their name (*Apple* and *Apple Inc.*). Some companies consist of multiple departments with different goals, but with the same name, or have departments in multiple cities. Occasionally, companies that are unrelated go by the same name, incidentally *Apple (Inc.)* 'Electronics company') and *Apple* (Recordings) 'Record company established by The Beatles'). The context in which a company name appears is often crucial to understand which company or department is referred to (Spina et al., 2011, 2013). Disambiguating between these company names is called *company name disambiguation* and can be approached with Entity Linking.

The companies in a country are generally registered by a Chamber of Commerce of the country. This organisation records some data about each company, usually at least its name, a description of its activities and a unique number. For company name disambiguation, it is useful to refer to companies by their unique number, to overcome name disambiguation. The description of activities could be used to differentiate between two companies with the same name.

Approaching company name disambiguation with Entity Linking would consist of resolving detected company names to their unique number in the register of the Chamber of Commerce.

1.3 Problem definition

A use-case for which the Entity Linking system proposed in this thesis can be found at Brainial, a company that creates software to facilitate tender procedures. A tender procedure is known as the process in which an organization is in need of goods or services and invites other parties to submit or propose a bid to provide these goods or services. In order to aid organizations to select the best bidding party, Brainial provides a service to retrieve current information about companies. One facet of this service is finding news articles about companies and linking these companies to entries in a database with additional information about the companies. At the moment, Brainial's method for this consists of comparing the name of the mentioned company with the company names in the database. The company name in the database, that is most similar to the mentioned company, is selected. As some companies have the same name, and other companies can be known under several different names, this approach can be faulty. A way to improve it, would be to use the context of the news article in which the company name appeared to decide which company to select. The goal of this thesis is to create an Entity Linking system that can link a company mention in a news article to a unique number in the company reference set, while making use of the context of the news article and the descriptions of the fields of specialization of the reference companies.The parts of this project ultimately lead up to the question:

Can Entity Linking successfully be used to perform company name disambiguation?

The contributions of this work include:

1. an exploration of using Entity Linking for company disambiguation in Dutch news articles.
2. the creation of a dataset with Dutch news articles annotated with company mentions and their corresponding KvK-numbers.

1.4 Terminology

In order to avoid any confusion with regards to the terminology used in this report, the most important terms are explained here.

- *mention*: name of a company that is written in a news article. The system will try to find the correct company number for each mention. A *mention* is referred to as an *alias* in the context of a Knowledge Base.
- *entity*: reference entity with a unique number. The system will be trained to link mentions of companies to company entities.
- *candidate*: company entity that a company mention potentially refers to. Some mentions have multiple candidates, others have only one.
- *Knowledge Base*: the set of unique company entities with additional information about them.

1.5 Outline

The remainder of this work is organised as follows: in Section 2, related work is discussed to provide some context for this project, followed by Section 3, which explains the methodology. Section 4 shows the results and in Section 5, the conclusion, discussion and future work are presented.

Chapter 2

2. Related Work

This chapter provides an overview of related work on Entity Linking, highlighting the more traditional and state-of-the-art approaches, and Entity Linking for Dutch data.

2.1 Named Entity Linking

As stated in the introduction, Named Entity Linking is the task of linking mentions of entities in texts to entities Knowledge Base. The Knowledge Base contains the reference entities and is either manually created to contain entities of relevance or it is an existing set, such as all entities on Wikipedia. One can divide Entity Linking into three stages: Mention Detection (MD): detecting the mentions of entities that need to be linked to entities in the Knowledge Base, Candidate Generation (CG): selecting a subset of entities the mention could possibly refer to and Entity Disambiguation (ED): deciding to which of the entities in the Knowledge Base a mention refers to. In some studies the Candidate Generation phase is not executed and all entities in the Knowledge Base are candidates for all mentions (Broscheit, 2020). Named Entity Linking is closely related to the task of Word-Sense Disambiguation, in which the use of a polysemous word is resolved to a specific meaning. The lexical database WordNet (Miller, 1995) is mainly used as a Knowledge Base for this task (Navigli, 2009).

2.2 Pipeline methods versus end-to-end methods

Common approaches consist of finding mentions to be linked in the Mention Detection stage, generating candidates for these mentions and disambiguating between the candidates. Traditional methods tackle MD and ED separately, making use of existing Named Entity Recognition systems for MD, while developing a custom system for ED. This can be referred to as a *pipeline architecture*, as mentioned in Finkel et al. (2005). It is called a pipeline architecture because systems that make use of this execute the mentioned phases of Entity Linking with (possibly different) models and in successive order. A positive side of pipeline architectures is that the researcher can make use of existing systems and combine them into an Entity Linking model. For example: a trained Named Entity Recognition system for the Mention Detection. A downside of pipeline architectures is that they allow for a propagation of errors (Kolitsas et al., 2018; Broscheit, 2020). This means that errors that are made in the one phase, are passed on to the next phase. For example, if an entity mention is not correctly recognised during the Mention Detection phase, because it contains extra words or only a portion of the correct mention, this erroneous mention will be passed on to the Candidate Generation. This may cause additional errors in the Candidate Generation phase, since it will generate candidates for the mentions it receives. An example of an error in the Candidate Generation phase is when the correct entity is not among the generated candidates. Then, the mention can never linked to the correct entity. Errors in the Candidate Generation phase can still occur when the Mention Detection was executed without errors.

A possible solution to overcome the problem of error propagation is to create an end-to-end system (Guo et al., 2013; Martins et al., 2019; Kolitsas et al., 2018; Broscheit, 2020; Durrett & Klein, 2014; Nguyen et al., 2016). This type of system will be trained to detect mentions, possibly to generate candidates and to link the mention to an entity. The input for these kind of systems is generally a document, and the output is a list of mention-entity pairs (Broscheit, 2020).

2.3 Approaches

This section will highlight some popular or famous approaches for Entity Linking.

2.3.1 Wikification

Many approaches use Wikipedia or Wikidata as a Knowledge Base and create systems to map entity mentions to pages on Wikipedia (Mihalcea & Csomai, 2007; Hachey et al., 2013; Rao et al., 2013; Qureshi et al., 2014; Tsai & Roth, 2016; Kolitsas et al., 2018). This approach results in a very general system that can be applicable for many use-cases and can easily be compared to other systems that are trained on Wikipedia or Wikidata. Wikidata[1] is a more structured version of Wikipedia. It links the same entities in different languages, labels all its entities with unique identifiers and provides short descriptions of them. This makes Wikidata suitable as a Knowledge Base for Entity Linking systems). One system for Wikification was developed by Rao et al. (2013). They approach the task as a ranking problem. First, a number of candidates is generated by several string comparisons, such as KB entities that are exact matches to the mention, KB entities that are wholly contained in or contain the mention, abbreviations of the mention, KB entities that are an exact match of a pre-computed alias list of the mention and KB entities that have a strong skip bigram Dice score or Hamming distance to the mention. A Support Vector Machine is provided with a feature vector representing the mention and each of the candidates from the KB and it will be trained to produce a score for each mention-candidate pair. An unspecified supervised machine learning ranker creates an ordered set of the candidates and is trained to rank the correct entity as the highest ranked candidate. The order of the other ranked candidates is not relevant. The feature vector that represented the mention consisted of a total of 26,569 features that includes combinations of the 200 base features (the SVM is not able to combine features as it is trained with a linear kernel, so combinations of features need to be provided separately). Among the features were name variant features (how similar the mention and the candidate are), Wikipedia features (candidate page length, number of links), popularity features and features about the context of the mention (other named entities that were present, cosine similarity of TF-IDF weighting between the context and the Wikipedia page of the candidate). The system was trained on 1300

[1] https://www.wikidata.org/

linked mentions and produced a maximum accuracy of 0.6639 on the Text Analysis Coreference 2009 (TAC-2009) evaluation set (McNamee & Dang, 2009) (only non-NIL mentions). The best system submitted to TAC-2009 reached an F-score of 0.7725 in this category and the median F-score of all submitted systems was 0.6352, so the system of Rao et al. (2013) performed better than most systems.

A well-known open source Entity Linking system is DBpedia Spotlight (Mendes et al., 2011), that can link textual mentions to entities DBpedia (Bizer et al., 2009). DBpedia is a Knowledge Base that contains structured data from Wikipedia. This includes globally unique identifiers for entries, relationships between entries and classification of entries into hierarchies. This structure facilitates for example linking a CEO to a company or connecting a product to its creator. DBpedia Spotlight was trained to disambiguate over 272 classes of mentions (persons, cities, companies). The system is rule-based and follows the traditional pipeline approach. First, existing information is exploited on DBpedia to create a lexicon. This information includes known mappings from entities to multiple different mentions and connections from multiple entities to an ambiguous mention. It uses *redirects* to other DBpedia pages and the *disambiguation pages*, that show lists of entities that a specific mention could refer to. In the same way, information from Wikipedia is included. Based on this information, prior probabilities (how often an entity was the correct link for a mention) are added, if available. Candidates for mentions are selected from the described lexicon. For disambiguation between the candidates, Mendes et al. (2011) they compute the cosine similarity between the TF-IDF vector of the paragraph surrounding the mention and the DBpedia page of each of the candidates. Instead of using all DBpedia pages for the Inverse Document Frequency, they use just the pages of the candidates, to ensure only terms that are important for disambiguation between only the candidates of a mention are selected. These cosine similarities are ranked and the highest ranked candidate is selected as the link for a mention. The authors annotated a test set with data from New York Times articles and compared the results of their own system with a number of other systems. DBpedia Spotlight received an F-score of 0.560 and was only beaten by The Wiki Machine[2], that received an F-score of 0.595. A number of other authors made use of DBpedia Spotlight in their systems (Olieman et al., 2014), (Bryl et al., 2015). Another ranking approach was done by Sil & Yates (2013), which is comparable to DBpedia Spotlight and will not be highlighted here.

For other cases, it may not be necessary to develop such a general system, because the Entity Linking is needed for a smaller domain. The main difference between domain-specific systems and general systems like those that use Wikidata as a Knowledge base, is that the Wikidata-based system could theoretically be used to link mentions in any kind of text to any kind of entity. For the domain-specific systems, the mentions to be linked usually have to meet certain demands, and the list of entities to link the mentions to is predefined and finite. An example of such a domain specific system is Hendriks et al. (2021). They created a system to link names of Dutch sailors from the 1600-1800's to names from notary records from the Dutch archives. This system will be described in more detail below.

Giles et al. (2005) applied an unsupervised K-way clustering method to disambiguate scientific authors in citations. Authors are generally cited with their first name initial and full last name, such as 'J. Lee.'. This makes room for ambiguity, as authors with different first names could be cited the same. Their Knowledge Base contained 14

[2] http://thewikimachine.fbk.eu

different author entities with their full names, to disambiguate the citations of authors to. They used three features: the co-authors that appear in the citation, the title of the paper and the publication organ, which were represented in a one-hot encoding with a weighted value. They experimented with the weight being a TF-IDF value and a Normalized TF value. The mentions in the dataset were clustered into 14 clusters with K-way spectral clustering and they achieved a maximum accuracy of 64.7%, averaged on all 14 authors. Other unsupervised Entity Linking methods were developed by Finkel et al. (2005) and Dalton & Dietz (2013).

2.3.2 State-of-the-art models

A recent end-to-end model was developed by Kolitsas et al. (2018). The neural-based system makes use of embeddings, without handcrafted features, to link mentions to entries of Wikipedia 2014. The input of the system is a document and the output is a list of mentions that occur in the document and the entity they are linked to. Their model consists of three feed-forward neural networks, with bidirectional LSTM's to encode the input of the system. The document that is inputted is represented in a one-hot encoding, with words from a dictionary. A bidirectional LSTM layer transforms these document vectors into context aware embeddings. For each span of the document, that could possibly be a mention to be linked, a set of candidate entities is added. These candidates are retrieved by Wikipedia hyperlinks and other dictionaries and are represented with pre-trained embeddings (built by Ganea & Hofmann (2017)). It is important to note that the Candidate Generation phase of Entity Linking is thus not included in this end-to-end model. Each candidate embedding is combined with prior probabilities and the embedding of the possible mention, and fed to a shallow feed-forward network to output a local score. As a final step, the authors add another feed-forward network to compute a global score, taking into account other mentions that occur within the same document. On a test set of 4,485 mentions, the created system received a macro F-score of 0.866 and a micro F-score of 0.894. These results are the highest when compared to other systems on the Gerbil platform (Usbeck et al., 2015; Röder et al., 2018), which is a general framework for benchmarking Entity Linking and other models. All systems were evaluated on the CoNLL-AIDA development and test sets (Hoffart et al., 2011).

Broscheit (2020) fine-tuned BERT (Devlin et al., 2018) to perform Entity Linking. They trained several variations of their model, but the base model will be highlighted here. As training data, they use the text on Wikipedia pages and the mentions on those pages that are associated with an internal link to another Wikipedia page. The titles of the pages that these mentions link to are viewed as the entities they refer to. They calculate prior probabilities for each mention-entity pair. The problem is viewed as a multi-class classification problem where the classes are the unique entities on Wikipedia, as they do not generate candidates for the mentions. The model was trained on 8.8 million mention-entity links for 4 epochs. The final result was an F-score of 0.828 on previously mentioned CoNLL-AIDA test set. Note that the system of Broscheit (2020) did not outperform the system of Kolitsas et al. (2018).

2.4 Named Entity Linking in Dutch

Not many systems for Entity Linking on Dutch data have been developed, that is why they are highlighted separately from the rest in this section. Hopefully, this gives a clear overview of what has been done for Dutch and makes it possible to grasp the position of the system proposed in this paper in the Dutch framework.

van Veen et al. (2016) developed a system to link named entities in Dutch historical news papers to Wikidata entries, to improve retrieval of the relevant papers as well as retrieving relevant information from them. They performed Named Entity Recognition on the newspapers with the Stanford Named Entity Recognizer (Finkel et al., 2005) and compared a rule-based approach for Entity Linking, grounded in various forms of string matching, to a machine learning approach, using Support Vector Machines and a set of features. These included character-level features to compare strings and context-level features. On a test set of 349 named entity mentions, the SVM approach outperformed the rule-based approach with with an accuracy of 0.831 compared to a score of 0.745. The results of their system were used to create an enrichment record for each mentioned entity, that stored the relevant Wikidata links. All these enrichment records together can be seen as a Knowledge Base, as described above. Although this approach is focused on historical news papers specifically, it can still be considered a general approach.

Another Entity Linking system for Dutch was developed by Hendriks et al. (2021) to link Dutch employees from the historical East-India Company to people from old notary records. This can be viewed as a domain-specific approach, because it focuses specifically on a subset of entities. In this case, the entities are certain people from a specific time in history. They compare Named Entity Recognition from spaCy (Honnibal et al., 2020) to BERTje (de Vries et al., 2019) and use Dedupe (Gregg & Eder, 2019) (a machine learning system for string matching) for linking the entities. Their system views the task as binary classification: given a mention-entity pair, does the mention refer to this entity? Their test set consisted of an unspecified representative number of mentions, featuring matches and non-matches, on which the F-score of their system was 0.846.

2.5 Company Name Disambiguation

A few studies have been done on company name disambiguation in Tweets. Some company names have another common meaning that refers to something in the real world that is not a company (such as *Apple*, the company, and *apple*, the fruit, or *Amazon*, the company and *the Amazon*, the river). It can be useful to know whether these ambiguous mentions actually refer to the company or not. Since Tweets do not contain many words, using the context will not always be helpful. Spina et al. (2013) created a binary classifier that, given a Tweet and an ambiguous company mention, predicts whether the mention in the Tweet actually refers to the company or not. Their approach relies on identifying *filter keywords*, whose presence in a Tweet reflect whether or not the mention refers to the company. Zhang et al. (2012) have a similar goal, but instead of identifying keywords, they use web resources to extract features that hold information about the companies to be disambiguated. They then train a number of machine learning methods (a.o. Naive Bayes and Support Vector Machines) to classify unseen Tweets. These approaches mostly contribute to the Mention Detection stage of

Entity Linking.

To our knowledge, no recent system has been developed to link mentions of company names to entities in another database. As many companies have a Wikipedia entry, company disambiguation is also partly covered in systems that make use of Wikipedia as a Knowledge Base (Rao et al., 2013), but it is usually not the main focus of a paper. The fact that not all Dutch companies are covered by Wikipedia underlines the need for a system that uses a custom Knowledge Base.

Chapter 3

Methodology

This chapter contains a description of the methodology that was used to create and train the Entity Linker. It includes a description of the data, the process of annotation to create a dataset that can be used for training, developing and evaluating the system created in this project. This section also features an explanation of the architecture of the created model and baseline systems.

It is important to note that the Entity Linking system will be designed to disambiguate between multiple (given) entity candidates, leaving the Mention Detection and Candidate Generation steps out of its scope. The mentions to be linked will be presented to the system during evaluation, as well as the candidates, that will be generated in the same way as the candidates presented in the annotation task. The system will not be able to predict the NIL-value for a mention. Therefore, mentions with no correct candidates will be excluded from the data presented to the system during training and evaluation. Mentions with only one candidate will also not be included, as disambiguation between candidates will not be possible for these mentions.

3.1 Data

The data used in this research consists of newspaper articles about companies and a database with their unique identifiers and additional information about the company entities in the Knowledge base. The data is provided by Brainial.

3.1.1 Web-scraped news articles

The news data provided by Brainial consists of news articles about business and companies that were scraped from Google News between 07/01/2021 and 09/04/2021. A number of queries was generated with topics retrieved from NEN, the Dutch Norm Institute. NEN administers norms and standards for companies regarding a large number of areas in the industry. The norms are aimed at companies and the topics in the queries were derived from these norms, which increases the chance that the web-scraped news articles with these topics feature companies. Examples of topics are: *Genetic Modification*, *Social Responsibility* and *Glass*. See NEN's website for a full overview of the norms and topics[1]. Included in the scraped news data were the titles, full texts and sources of the articles.

[1] https://www.nen.nl/

3.1.2 Company database

In addition to the news scraped data, Brainial provided a database with information about 8,590 Dutch companies. For each company, a unique number, main name, alternative names and a description of the activities of the companies were included. This company description is represented with an SBI-code that categorizes the company based on its main activity and an SBI-code description, which is a compact description of the activity.

Unique identifiers

The Dutch Chamber of Commerce is called the Kamer van Koophandel (KvK). This is the official Dutch administrative body for businesses. All companies in the Netherlands have a unique number provided by the KvK, which is referred to as KvK-number. The KvK records several types of information about companies, such as a short description, city the company is based in and the number of employees. Different departments of the same company sometimes have their own KvK-number. These KvK-numbers will be used as unique numbers for the companies in the Knowledge Base.

Alternative company names

The alternative company names were retrieved by Brainial from the Kamer van Koophandel. Example of alternative names are deprecated names of companies and abbreviations of the main name (such as Kamer van Koophandel (KvK) and Centraal Bureau voor Statistiek (CBS)).

SBI-code description

SBI is an abbreviation for 'Standaard Bedrijfsindeling' (Standard Company Division) and is used to divide companies into categories, based on their main activity, with unique codes per category. All SBI-codes are accompanied by a textual description that can be found on the website of the Dutch Central Bureau of Statistics[2]. In total, 943 SBI-codes exist. The SBI-code descriptions will be treated as descriptions of the companies in the database and will be used to construct a Knowledge Base. This Knowledge Base will be available for the system during training to provide the descriptions of the company entities that the system has to link company mentions to.

3.1.3 Preprocessing

Both the news article dataset and company database were preprocessed, respectively to be used as training, development and test data and to construct a Knowledge Base for the system. The preprocessing steps for each of the two datasets are described below.

Preprocessing of news-scraped data

At first, three preprocessing steps were executed:

[2] https://www.cbs.nl/nl-nl/onze-diensten/methoden/classificaties/activiteiten/sbi-2008-standaard-bedrijfsindeling-2008

- The first paragraph of the article was extracted by splitting the article text on double newlines. When the first component was the same as the title, the second component was saved as the first paragraph. In the other cases, the first component was saved as the first paragraph.

- Some full texts of articles did not include the title, so in those cases, the title was added to the full text.

- The full text was stripped of double spaces and newlines.

Named Entity Recognition was performed on the articles to extract the company names. spaCy (Honnibal et al., 2020) provides a NER-system for Dutch, which was additionally trained by Brainial to improve its performance on company name detection. As NER is out of scope for this project, the process of improving the NER-system will not be clarified here. The NER system of spaCy categorizes the recognised entities into a type, such as person, location and organization. Since the focus of this project is on company names, only the entities of type *ORG* and *NORP* were extracted from the news articles, as these entity types generally refer to the companies that have a KvK-number. *ORG* entities always refer to 'organizations' and *NORP* refers to 'nationalities or religious or political groups'. These types of Named Entities derive from the corpus the Dutch spaCy model was trained on: Named Entity annotations on the Lassy Small corpus Van Noord et al. (2013) by NLP Town[3]. Political groups can be viewed as companies and tend to have a KvK-number, which is why this entity type was included in the data. Only articles that did contain one or more named entities of the type *ORG* or *NORP* were included in the data. After preprocessing, the news-scraped data included the full texts of the articles, the titles, the first paragraphs and URL's to the source of the articles.

Preprocessing of company database

As a first preprocessing step, the main name and alternative names of the company entries were merged into a list of all names. Providing all known variations of a company name will be useful for generating candidate companies for company mentions in the articles, since all known names will be linked to one entry. For some company mentions, the SBI-code and SBI-code description were not available. The SBI-code descriptions are part of the input to the system, so it would not be possible to link a company mention to a company entity that does not have a SBI-code description. Therefore, only company entries with such a description were included in the company database. Some KvK-numbers contained seven instead of eight numbers, which means the numbers are deprecated, because currently, all KvK-numbers contain eight numbers. By adding a zero at the beginning of the number it becomes valid again. This was applied to all KvK-numbers of seven numbers. Finally, the KvK-number, main company name, all company names, SBI-code and SBI-code description were saved.

3.1.4 Annotation

The company mentions in the news articles were manually annotated with the KvK-numbers they refer to, or were excluded from the data when it was not possible to link

[3] https://nlp.town/

them to a company entity from the Knowledge Base. Since the designed system will be trained to disambiguate between candidates, only mentions with multiple candidates were annotated. Mentions with no candidates cannot be linked to an entity company and will thus be disregarded.

Candidate generation

The total number of KvK-numbers within the scope of this research is 8,590, so to compare each company mention with all 8,590 numbers can take a long time. In order to facilitate the annotation process, candidate KvK-numbers were generated for each company mention, by comparing the company mention to the names and alternative names of the companies in the database of KvK-numbers. An n-gram based implementation of fuzzy matching was used to do this. All company names were split into trigrams of which a TF-IDF (term frequency-inverse document frequency) vector was computed, effectively treating the trigrams as terms and the company names as documents. The similarities between the company mention's TF-IDF vector and TF-IDF vector of each alternative name of the company entities was computed with a fast implementation of cosine similarity. The company entities with the highest cosine similarities, above a threshold of 80%, were selected as candidates for the company mention, with a maximum of 5 candidates per mention. This number was chosen to facilitate the annotating process by limiting the number of candidates to be compared by the annotators. The approach of generating candidates with fuzzy matching to facilitate annotation was inspired by Hendriks et al. (2021). A short analysis of the number of candidates per mention revealed that most company mentions only had 2 candidates, a few had 3 or 4 candidates and none of the mentions had 5 candidates. This suggests that for each mention, all possible candidates were generated and limiting the number of candidates to 5 did not influence the system in any way.

Task

The annotators were presented with the company mention to be annotated, the title of the news article, the first paragraph of the news article and a portion of the article in which the company mention occurred. If there were multiple occurrences of the same company mention in the text shown, the article portion included only the first occurrence of the mention. Each annotation sample also included a URL to the full article on its original website. In addition, annotators were shown a number of options to choose from, which will be explained below. It was decided to not immediately show the full article during annotation to ease the task for the annotators. By presenting less text on the screen, reading the annotation sample would take less time and annotators would not have to scroll down to see the options. Different company mentions that occurred in the same news article were presented separately but consecutively. For each annotation sample, annotators could select one of the following options:

- **A candidate from the candidate list.**
  Each candidate was presented with its base name, its SBI-code description and its KvK-number. The task was about selecting the correct candidate based on the news article.

- **Mention is not a company/company is not recognised correctly.**
  This option was aimed at samples in which it was clear that the proposed company

mention was not actually a mention of a company. For example, the word *vijf* in Dutch is the ordinal number *five*, but it is also a furniture company. When it was clear from the article that the ordinal number was meant, the option described here had to be chosen. Other scenarios in for which this option was meant were samples with proposed company mentions that contained extra words from the article, such as *Naast Itho Daalderop* (in which only *Itho Daalderop* would have been the correct mention) and proposed mentions that did not consist of the full company name, such as *ABN* (in which *ABN AMRO* would have been the correct mention). These problems were caused by the Named Entity Recognizer of spaCy.

- **Correct entity is not in candidate list.**
  This option was intended for samples in which none of the presented candidates seemed to be the correct candidate for the proposed mention. It was assumed that the correct candidate was then not included in the Knowledge Base and it would not be possible to link the mention to a KvK-number.

- **Not enough context.**
  This option was meant for samples in which the article did not provide enough context to decide which candidate would be correct. This usually occurred in articles that listed a large number of companies, for example companies that were all affected by an economic crisis. These companies could be very diverse and if the article did not include context about the individual companies, it could be impossible to select a candidate.

A few articles were written in English but contained Dutch company mentions. These articles were rejected by the annotators, as the system will only be trained on Dutch articles. The full annotation guidelines can be found in Appendix A.

Inter-annotator Agreement

In total, two annotators annotated 3752 company mentions. The number of samples each of them annotated is shown in table 3.1.

Table 3.1: Number of annotations per annotator

Annotator	**Annotations**	**Used as data**
Annotator A	2106	1717
Annotator B	1646	1403
Total	3752	3287

In addition, 370 samples were annotated by both annotators to calculate inter-annotator agreement and the Cohen's Kappa. 281 mentions were annotated the same, resulting in an inter-annotator agreement of 0.759, with a Cohen's Kappa of 0.732. These samples included company mentions that were not linked to an entity in the Knowledge Base, as described in the annotation task. Of the 370 samples, 187 were linked to a KvK-number by both annotators and 167 were linked to the same KvK-number. The inter-annotator agreement of these samples was 0.893 and the Cohen's Kappa was 0.892. The 167 samples that were annotated the same by both annotators were added to the data, since the data only includes mentions that were linked to an entity from the Knowledge Base. This resulted in the final dataset consisting of 3287 mentions linked to a KvK-number.

3.2 Architecture

This section describes the architectures of both the system that has been created for this project and the baselines to be compared to the performance of the system.

3.2.1 Entity Linking component in spaCy

spaCy (Honnibal et al., 2020) provides a custom Entity Linking module that can be trained with self-provided data. It needs a Knowledge Base that stores information about the aliases under which entities may appear and description vectors about these entities. The module can be trained with annotated data and added to the processing pipeline, to be used on unseen data in the future. The module will be trained to predict how likely it is that the description of a candidate entity is similar to the context of a mention. The candidate with the highest prediction will be selected as the correct candidate. Only company names that are recognised by the Named Entity Recognition component from spaCy are considered and within this project, only Named Entities of type *ORG* and *NORP*. After training, the system can be queried to predict entities for mentions that are recognised by spaCy's NER. If the exact mention exists in the Knowledge Base, it will make a prediction for each of the candidate entities, with regards to the text in which the mention occurred, and return the candidate with the highest probability. When there is only one candidate, this will automatically be returned. Company mentions that do not exist as an alias in the Knowledge Base cannot be subject to a prediction of the system and will receive the NIL (not-in-lexicon) label. The decision to utilise spaCy for the Entity Linking system was made because of its speed and easy integration into Brainial's environment, as they make use of spaCy for other components of their activities.

Knowledge Base architecture

Prior to training, a Knowledge Base for the Entity Linking module was constructed. The Knowledge Base consists of two elements:

- **The entities with vector representations of their SBI-code descriptions.** First, vector representations of the SBI-code descriptions were obtained from the large spaCy model for Dutch (*nl core news lg*). This model was trained on the Dutch Wikipedia and web-crawled news data and contains word vectors of 300 dimensions. The vector representations were constructed by averaging the embeddings of the words in the descriptions. Then, the embeddings were encoded into a vector with a dimension of 64 with a pretrained encoder-decoder system from spaCy. The resulting, lower dimension vector was stored in the Knowledge Base with the entity it describes.

- **Aliases of mentions with their candidates**
  The second element of the Knowledge Base contains all company mentions from the data (training, development and test data) and stores their candidates. The candidates were generated in the same way as the candidates that were generated for the annotation phase: the company mention was compared to all names of the entities in the list of entities, and a maximum of five candidates that were 80% or more similar to the company mention were selected (see subsection **Candidate generation** in **Annotation** for more details). Within this candidate list, each candidate was accompanied by a prior probability: a number between 0 and 1 stating how often this candidate was the correct one for this mention. The probabilities were calculated by counting how often the candidate was correct for a certain mention and dividing this number by the number of times the mention occurred in the annotated data.

Structure of Entity Linker

The Entity Linker from spaCy is a feed forward neural network, originating from a machine learning library called Thinc[4] with an input layer of size 243, and includes a mean pooling layer and an output layer with linear regression activation to output a number between 0 and 1. Every sample that is fed to the Entity Linker is a feature vector of four elements:

1. *The context of the mention.*
   The context is the text the mention appeared in. Within the Entity Linker, this could either be just the sentence in which the mention occurred or a specified number of sentences around the sentence of the mention. The text of the context will be represented with averaged word embeddings of size 300 from the Dutch spaCy model *nl core news lg*. The size of this vector was reduced to 128 with a Convolutional Neural Network consisting of 4 convolutional layers. The final vector of size 128 represents the context of the mention. It was decided that the system will be trained with two variations of this input: 1. Only the sentence in which the mention occurred and 2. All sentences in the article. The longest article in the training data contained 1427 sentences and the shortest article contained 5 sentences.

2. *The type of Named Entity of the company mention.*
   This will be encoded in a one-hot vector of size 18, since this is the number of different Named Entity types that exist in the spaCy's large model for Dutch. The model will only regard mentions of type *ORG* and *NORP*.

3. *The description vector of the candidate entity.*
   The descriptions of the candidate entities (as described in the **Knowledge Base Architecture** section) are represented with a 96-size vector.

[4] https://thinc.ai/

4. *The prior probability of the candidate entity.*
   The prior probabilities of the candidate entities (as described in the **Knowledge Base Architecture** section) are floating point numbers between 0 and 1 stating how often the concerned candidate was the correct one.

The labels that accompany the input samples are values of 1 or 0, respectively when the candidate was the correct one for the mention in this context, and when the candidate was not.

3.2.2 Baseline systems

Three baselines were constructed to be able to compare the results of the trained Entity Linking system to other systems and to measure the difficulty of the task. Each will be described in this section.

Majority Baseline

It was decided to compare the results of the system to a majority baseline. A prediction for a mention was made by retrieving the candidates and their prior probabilities from the Knowledge Base of the Entity Linker, and selecting the candidate with the highest prior probability. In this way, the context of the company mention was disregarded.

Brainial's baseline

Brainial's current approach for the task consists of fuzzy matching the company mentions in the news articles to the companies in the database with KvK-numbers and company names. The company mention is compared to all names and alternative names of each company in the company database, and the KvK-number of the company name that has the highest similarity to the company mention in the news article will be returned, if their similarity is 74% or higher. The similarities are calculated as described in section **Annotation**, subsection **Candidate generation**. When the highest similarity between a company mention and a company from the database is less than 74%, none of the companies will be returned, as it is in that case likely that the company names differ too much to refer to the same entity. The company mention will then receive the NIL label. The value of 74% has been determined by Brainial after a number of experiments.

Brainial's baseline with context comparison

A third baseline was designed, that incorporated the context of the company mention (the news article it appeared in) and the SBI-descriptions of the candidates. Instead of selecting the KB entity with the highest similarity, a maximum of 5 candidates with the highest similarity were generated, with a threshold of 80% similarity. To compare the context of the news article with the SBI-descriptions of the candidates, a document vector of the new article was constructed by averaging the word embeddings of the word in the article. In the same fashion, document vectors for each of the candidates' SBI-descriptions were retrieved. The word embeddings from spaCy's large model for Dutch (*nl core news lg*) were used. The cosine similarity of each SBI-description vector and the document vector of the news article was computed, and the candidate with the highest cosine similarity was selected.

Table 3.2: Organization of the data

Data	**Company mentions**	**Articles**
Training data	1972	1675
Development data	658	623
Test data	658	619

3.3 Training

60% of the total number of accepted company mentions was used as training data for the Entity Linking system. Each sample of the data consisted of the company mention and the full article it appeared in. Company mentions that occurred in the same article were treated individually and were each provided with the same article. The system that was trained on just the sentence in which the mention occurred received only that sentence. The system received the training samples in mini-batches of increasing size using compounding, starting with a batch-size of 4 and gradually multiplying the batch-size with 1.001 to reach the maximum batch-size of 32. Compounding has been proven to be effective by Smith et al. (2017). The system was trained for 500 iterations. A dropout value of 0.2 was set to prevent overfitting. The system was configured to include the prior probabilities from the Knowledge Base and to only consider mentions of Named Entity type *ORG* and *NORP*. As mentioned before, the model was trained twice, once with the single sentence in which a mention occurred as context and once with the full article as context. The system will be intermediately tested on a development set. The total data consisted of 3456 company mentions that were linked to an entity in the Knowledge Base, occurring in 2503 news articles. 60% of the data will used as training data, 20% will be used as development data and another 20% will be used as test data. Table 3.2 shows the exact number of mentions in the training, development and test data.

3.4 Evaluation

Both the trained system and the three baseline systems will be evaluated on 20% of the total data, featuring 692 company mentions. The systems receive as input pairs of full articles and company mentions that appear in those articles and they have to predict the Knowledge Base entity for each mention. All systems will be evaluated on micro F-score, recall and precision.

Chapter 4

Results

4.1 Results on the test set

In total, five systems have been evaluated on the test dataset, of which three baseline systems: a majority baseline, Brainial's baseline (*Brainial's baseline*), Brainial's baseline with context comparison (*Brainial's baseline+context*). Two versions of the Entity Linking system that has been developed will be evaluated: one in which the system received the full article as context for a company mention (*EL (full article)*) and one in which the system received only the sentence in which the company mention occurred as context (*EL (one sentence)*. The results are presented in table 4.1.

Table 4.1: Micro results of baselines and trained models on the test set.

System	F-score	Recall	Precision
Majority Baseline	0.929	**0.946**	0.924
Brainial baseline	0.629	0.647	0.623
Brainial baseline+context	0.647	0.627	0.806
EL (full article)	0.931	0.939	0.927
EL (one sentence)	**0.935**	0.941	**0.931**

The highest F-score and precision on the test are achieved by the *EL (one sentence)* system, though the results are very close to the *majority baseline*, that achieves the highest recall. The *EL (full article)* system performs worse on the test set than *EL (one sentence)* and the *majority baseline*, but better than the Brainial baselines. From the two Brainial baselines, the *Brainial baseline* outperforms the baseline with context comparison (*Brainial baseline+context*). This last system is the only one with a precision value that is not comparable with the recall and F-score values, but much higher.

4.2 Error Analysis

This section describes the results of an error analysis that was performed on the output of the best performing system to get insight into the type of errors it makes and what may have caused them.

In total, the Entity Linking system that received only one sentence as context, made 31 errors on the test set, resulting in an error percentage of 5.0%. The errors can be categorized into groups.

There are a number of reasons why the system makes errors, which could either have to do with the SBI-descriptions or the context that that system received during training.

1. **SBI-description of the gold-label and of the prediction are the same.** For some samples, the SBI-code descriptions of the predicted label and the gold-label were identical, but the predicted KvK-number was different. The SBI-description is one of the four features that make up the input for the model, and when one of the vectors is the same for candidates, it should be harder for the system to learn what distinguishes the two and therefore, make the correct prediction.

2. **SBI-description of the gold-label is not informative.** The SBI-code description of the gold-labels and predictions could be insufficient in reflecting the companies main activity. There are a number of SBI-descriptions that are very broad and describe activities related to companies in general, related to laws and finances. These descriptions will commonly not be reflected in the context of the company mention. Examples are: *Financiële Holdings* ('Financial Holdings'), *Holdings (geen financiële)* ('Holdings (not financial') and *Lease van immateriële activa* ('Lease of intangible assets').

3. **The sentence in which the mention occurred did not provide enough context to disambiguate between the candidates.** Since the data was annotated while taking the full article into account, but the best performing system only took the sentence in which the sample occurred into account, some sentences did not provide enough context to disambiguate between candidates, while the full article did. This was the cause for a number of errors. Another example in which the context was not informative enough is when it lists a large number of companies without other information.

4. **Other.** Cases in which the error that was made cannot be placed in the categories above will be placed in this category.

Table 4.2 shows the distribution of the error types on the test data and each error type is accompanied with an example error.

Table 4.2: Percentage of errors and example error of each error type. For the gold-label and prediction, the predicted KvK-number is shown, followed by the SBI-code description connected to the KvK-number.

Error type	**Percentage**	**Example**
1	29%	Company mention: *Evides* Gold-label: **24388995** (*Winning en distributie van water*) Prediction: **24170650** (*Winning en distributie van water*) Sentence: *De verwachting is dat Evides in het voorjaar van 2022 het eerste water zuivert.*
2	16%	Company mention: *FrieslandCampina* Gold-label: **11057544** (*Financiële holdings*) Prediction: **1070163** (*Lease van niet-financiële immateriële activa*) Sentence: *Dens uit Helmond gaat met Heijmans bouwplaats op met emissieloos aggregaat.*
3	26%	Company mention: *Siemens* Gold-label: **53745175** (*Productie van elektriciteit door thermische, kern- en warmtekrachtcentrales*) Prediction: **27015771** (*Vervaardiging van communicatieapparatuur*) Sentence: *Siemens schrapt wereldwijd.*
4	29%	Company mention: *Eneco* Gold-label: **24242021** (*Productie van elektriciteit door thermische, kern- en warmtekrachtcentrales*) Prediction: **24296168** (*Handel in elektriciteit en in gas via leidingen*) Sentence: *Warmtebron Utrecht, een samenwerking tussen onderzoekers en ENECO, heeft Hoek Zuidstede in Nieuwegein gekozen als locatie om het onderzoek naar aardwarmte voort te zetten.*

Chapter 5

Concluding remarks

5.1 Conclusion

This thesis investigated the possibility of using Entity Linking to perform company name disambiguation. An Entity Linking system was developed to link company mentions in news articles to unique identifiers in a custom Knowledge Base, by disambiguating between candidates. As a part of this project, an annotated dataset of company mentions in news articles was created that can be used for evaluation of future systems. The created system greatly outperformed two baseline systems and reached a micro F-score of 0.946 on an evaluation set, but the scores are comparable to a third, majority, baseline. This suggests that the task at hand might not be as difficult as expected, provided there is annotated data. Approaching company name disambiguation with an Entity Linking system is useful, but a majority baseline yields the same results and is simpler.

5.2 Discussion

This section will discuss the results, the quality of the data and some drawbacks to the method of using the Entity Linking system of spaCy.

5.2.1 Discussion of results

The Entity Linking model that was developed in this study did not perform much better than the majority baseline that was constructed. The differences in F-score, recall and precision are almost negligible. Both of these systems do greatly outperform the two Brainial baselines, but the results suggest that a majority baseline suffices and it may not be necessary to train an Entity Linking system. This would mean that there is not that much ambiguity within company names and the context in which a company mention occurs has little influence on what entity should be linked to a mention. Even though some companies do have different departments with various goals and different KvK-numbers, it might just be that the companies mentioned in news articles mostly refer to the main branch of an organization. News articles are commonly written for the general public that might not know about all departments of a company and it would be clearer to just refer to the main branch. That being said, there are still a couple of interesting points to discuss with regards to the Entity Linking system.

The system performed better when it was trained on solely the sentence in which the company mention occurred than when it was trained on the whole article. The opposite was expected, since the full article contains more information and there is chance that the most important information is not included in the single sentence of the company mention. On the other hand, some company mentions are preceded by a descriptive term, such as *energiebedrijf Eneco* ('energy company Eneco') which could be enough context for correct disambiguation. The error analysis makes clear that the lack of sufficient context significantly contributes to the errors that were made (26%). Still, the number of errors was very small, so it does not seem to be a very big problem. A possible explanation is that the often large article that is transformed into a vector becomes very general because it consists of the averages of many words, while this is not the case for the shorter entity descriptions. The two vectors could become very different, even though the entity description would match the news article.

5.2.2 Data quality

The database of company entities and KvK-numbers was not complete; that is: it did not contain the KvK-numbers of all Dutch companies. It is not necessarily bad to evaluate the systems on a subset of all KvK-numbers for research purposes, but it does matter for the usefulness of the created system for Brainial. The database can be expanded in the future, but the system will also need extra annotations featuring the entities that were added.

The SBI-descriptions of the company entities were not all equally informative. For example, the SBI-code description *Organisatie-adviesbureaus* ('Organisational consultancies') is less detailed than *Beheer en exploitatie van transportnetten voor elektriciteit, aardgas en warm water* ('Management and operation of transmission grids for electricity, gas and hot water'). This could have influenced the systems ability to link a description to the context of a news article. In addition, some candidates for the same mention coincidentally had the same SBI-description. The system was trained to disambiguate between candidates based on the SBI-code description, so naturally it would not be possible to disambiguate between candidates that have the same description. This calls for using other types of descriptions for the company entities, such as a biography on their website or social media, or some information from their Wikipedia page. However, this information is not available for every company and would be more difficult to extract. It would be very time consuming to produce this for every company.

The list of alternative names that was available for most companies was also not always complete. For example, the abbreviation *KvK* for the *Kamer van Koophandel* was not included, so when the system encountered *KvK* as a company mention, the *Kamer van Koophandel* would not be proposed a a candidate, and as a result, this mention could not be linked to the correct entity, even though this entity was included in the Knowledge Base. The influence of this problem propagated to the annotation process as well because the mentions were annotated with the candidate KvK-number that seemed most likely to be correct. These candidates are retrieved in the same way as the system retrieves candidates for a company mention. So when the correct entity is not among the candidates, the mention can also not be annotated correctly. It would be interesting to see how the system performs on a test set of which the annotation was done independently of the systems candidate generation.

5.2.3 Drawbacks of spaCy's Entity Linker

Finally, there are three main drawbacks to the Entity Linker of spaCy.

1. First, the system is not able to predict the NIL-value when the correct entity for a mention is not in the Knowledge Base, but the mention still has candidate entities. When the correct entity for a mention is not included in the Knowledge Base, the correct prediction should be NIL. If one or more candidates can be generated for a mention, the Entity Linker will select one of those candidates, even though none of the candidates is correct. An example is the mention *Amazon*, referring to the American web-shop. It is not a Dutch company so it does not have a KvK-number and thus is not included as an entity in the Knowledge Base. The Knowledge Base does include a different company called *Amazon*, that is an actuarial and pension consultancy firm, located in the Netherlands. As these two company names are a 100% similar, the Dutch *Amazon* will be proposed as a candidate for the American *Amazon*. This is the only candidate for the mention *Amazon*, which means the system will always select this candidate, even though it was not trained to do this. The problem can be overcome by setting a probability threshold. The system predicts a probability value for each candidate and the candidate with the highest probability value is selected, even if the highest probability is relatively low. By setting a threshold, the NIL-value could be predicted in the case that the probabilities for all candidates are below the threshold. This might also make it possible to train the system to predict KvK-numbers for mentions with only one candidate. It would then either select the candidate or return the NIL-value. Sadly, it is currently not possible to implement this in the spaCy's Entity Linker, so creating this feature needs to be saved for future work.

2. Another drawback of the Entity Linker is the fact that it can only predict a KvK-number for mentions that are exact matches of aliases that already exist in the Knowledge Base. Company mentions that the system has not seen before automatically get the predicted the NIL-value, even if the correct entity exists in the Knowledge Base. Both of Brainials baselines can make predictions for values they have not seen before, which makes them more robust than the Entity Linking system. A possible improvement for this could be to select aliases from the Knowledge Base that are a close match to the mentions subject to prediction, for example with fuzzy matching. The candidates for these aliases could then be regarded as candidates for the mention, and a prediction could be made. This would only apply to mentions for which a close match exists in the Knowledge Base and it would nonetheless still not be possible to make predictions for mentions for which this is not the case.

3. A last issue that has been encountered in the functionality of spaCy's Entity Linking is the lack of full customization of the context that can be provided for the company mention during training. It can be informative to experiment with the amount and type of context the system receives for each mention. In news articles, the main topic of the article is usually mentioned in the title and/or at the beginning of the article. The rest of the article (especially when its lengthy) could make the context representation noisy, since it is the average of the individual word embeddings. The Entity Linker does make it possible to adjust the number of sentences around the mention, but it is not possible to customize the context

before feeding it to the system during training, as it works as follows: for every training sample, the Entity Linker performs NER on the context and checks if the mention in the sample is recognized in the context. A problem arises when the context is modified after its company mentions were annotated: as the context is different, the input for the NER-system changes and this might result in the case that the mention in the sample will not be recognised in the context. As a result, this sample could not be used as training data. Since the system is eventually trained to match an entity description to a context, this should theoretically not be a problem, but currently it practically is.

5.3 Future Work

This section describes the suggestions for future research.

A first suggestion would be to manually create a system that is similar to spaCy's Entity Linker, so that the drawbacks stated above could be avoided. This would make it possible to request the probabilities of the predictions and apply a threshold, and to generate candidates for company mentions the system has not encountered in the training data. It would also be possible to experiment with the context that is provided since the system does not have to actively recognise the mentions in the context anymore. This would resolve the three drawbacks of spaCy's Entity Linker that were mentioned in the discussion. It would then also be possible to use a different Named Entity Recognition model for the mention detection stage, for example that of BERTje (de Vries et al., 2019), that receives a higher F-score (0.883) than the score reported for spaCy's NER (0.77). These scores were computed on different test datasets, but it does indicate that BERTje outperforms spaCy's model. Using BERTje could then improve the Mention Detection stage and therefore the usefulness of the system.

It would also be wise to expand the company database that was used as Knowledge Base. Including more reference entities would result in more links between mentions and entities and more training data and therefore most likely improve the system.

Lastly, provided the resources would allow it, it could be possible to create an end-to-end system for company name linking, with an architecture inspired by the architectures of (Kolitsas et al., 2018) or (Broscheit, 2020). The annotated dataset that was created in this project could be used to start training the end-to-end system, but it would probably need many more annotations in order to reach state-of-the art performance.

Appendix A

Annotation Guidelines

A.1 Beschrijving van de taak

Het doel van deze annotatietaak is om een bedrijfsnaam dat voorkomt in een nieuwsartikel te koppelen aan een KvK-nummer. Bedrijven worden vaak bij verschillende namen genoemd, zoals *Apple* en *Apple Inc.* Soms kan er met dezelfde bedrijfsnaam 2 verschillende bedrijven met andere KvK-nummers worden bedoeld, zoals *Jansen (Loodgieters)* en *Jansen (Financieel Adviseurs)*. Voor elk bedrijf dat genoemd wordt in een nieuwsartikel worden één of meerdere KvK-nummers laten zien, met de basisnaam van dit bedrijf en de sector waarin het actief is. De taak van de annotator is om op basis van de context van het nieuwsartikel en extra informatie over de KvK-nummers waar de bedrijfsnaam naar zou kunnen verwijzen, het correcte KvK-nummer te selecteren. Het is ook mogelijk dat de bedrijfsnaam aan geen van de optionele KvK-nummers kan worden verbonden.

Een annotator krijgt de titel, de eerste paragraaf en 500 tekens van het artikel waarin de bedrijfsnaam voorkomt te zien.

Een KvK-nummer wordt getoond tezamen met de basisnaam van het bedrijf waar het nummer bij hoort en de SBI-code omschrijving, die aangeeft in welke sector het bedrijf actief is. Deze verzameling van gegevens wordt een *kandidaat* genoemd.

A.2 Hulpmiddelen

Tijdens het annoteren zijn er een aantal hulpmiddelen ter beschikking, die hier worden genoemd en toegelicht. Naast de link naar het volledige artikel zijn er voor elke kandidaat een aantal hulpmiddelen beschikbaar.

- **De basisnaam van het bedrijf.**
  Dit is de hoofdnaam van het bedrijf, zoals genoemd in de database die is gebruikt voor deze taak. De database bevat ook alternatieve namen, dus het kan voorkomen dat de bedrijfsnaam die genoemd wordt in het artikel niet direct overeen lijkt te komen met de basisnaam van de kandidaat. Het is dan nog wel mogelijk dat deze kandidaat de juiste is.

- **Link naar het KvK-nummer in het Handelsregister op de website van de Kamer van Koophandel.**
  Door op het KvK-nummer van een kandidaat te klikken, wordt de annotator doorwezen naar het nummer in het Handelsregister op de website van de Kamer

van Koophandel. Hier staat meer informatie over het bedrijf dat bij dit KvK-nummer hoort, zoals de basisnaam zoals die in het Handelsregister staat, een vestigingsadres en de website van het bedrijf. Soms zijn er meerdere resultaten in het Handelsregister met hetzelfde Kvk-nummer en vergelijkbare basisnamen. Wanneer een resultaat uit het Handelsregister dezelfde SBI-code omschrijving en KvK-nummer heeft, maar een andere basisnaam dan de kandidaat in Prodigy, en deze basisnaam lijkt meer op het bedrijf dat genoemd werd in het artikel, dan kan deze kandidaat nog steeds de juiste zijn, omdat het in deze taak draait om KvK-nummers.

- **De SBI-code omschrijving.**
  De omschrijving van de SBI-code geeft bondig aan wat de hoofdactiviteit is van een bedrijf. Deze omschrijving kan duidelijk maken bij welk KvK-nummer het genoemde bedrijf hoort, wanneer de omschrijving overeenkomt met de context van het artikel, of juist niet.

- **Toelichting op de SBI-code omschrijving.**
  Wanneer de SBI-code omschrijving onbekend is bij de annotator of niet genoeg informatie geeft, kan er op de omschrijving worden geklikt, waarna de annotator met een URL wordt doorgestuurd naar de website van het Centraal Bureau voor Statistiek (CBS) waar een toelichting op de desbetreffende SBI-code omschrijving staat.

- **Link naar het volledige artikel.**
  Wanneer het getoonde stuk van het nieuwsartikel niet voldoende context biedt om de juiste keuzeoptie te selecteren, kan het volledige artikel opgevraagd worden via de link om zo meer informatie te krijgen. Soms geeft de bron of website van het artikel al veel informatie over de sector van het genoemde bedrijf, indien deze vooral artikelen bevat over één sector, zoals economie of technologie. Sommige artikelen zijn onbeschikbaar geworden in de tijd tussen het verzamelen van de artikelen en het annoteren van de bedrijfsnamen en kunnen dus niet op de originele website worden weergegeven.

- **Extra informatie zoeken met Google**
  Het is toegestaan om extra informatie over alle kandidaten en het artikel te zoeken op Google. Soms wordt er in een niewsartikel verwezen naar de website van het genoemde bedrijf en is daarop informatie te vinden waardoor duidelijk wordt welke kandidaat de juiste is, door bijvoorbeeld de vestigingslocatie te vergelijken met de vestigingslocatie van het KvK-nummer in het Handelsregister.

A.3 Keuzemogelijkheden

Voor de taak kan er gekozen worden uit 5 of meer opties, afhankelijk van het aantal kandidaten van het genoemde bedrijf:

- **2 of meer kandidaten.**
  Deze optie van een kandidaat wordt gekozen wanneer de annotator zeker weet welke van de getoonde kandidaten overeenkomt met het bedrijf dat genoemd wordt in het nieuwsartikel, nadat er gebruik is gemaakt van de hulpmiddelen.

- **Bedrijf staat niet in de opties.**
  Deze optie wordt gekozen wanneer de getoonde kandidaat niet overeenkomt met het bedrijf dat genoegmd wordt in het nieuwsartikel, en deze dus verwijst naar een ander KvK-nummer dat niet in de opties staat.

- **Herkende 'bedrijf' is geen bedrijf.**
  Deze optie wordt gekozen wanneer de besdrijfsnaam die herkend is in het artikel eigenlijk geen bedrijfsnaam is. De oorzaak van deze fout ligt buiten het domein van deze annotatietaak en deze bedrijfsnamen worden niet meegenomen in het systeem waarvoor de annotaties gebruikt worden.

- **Niet genoeg context.**
  Wanneer het, na gebruik van de hulpmiddelen, onmogelijk is om te bepalen welke van de keuzeopties de juiste is, kan deze optie worden gekozen.

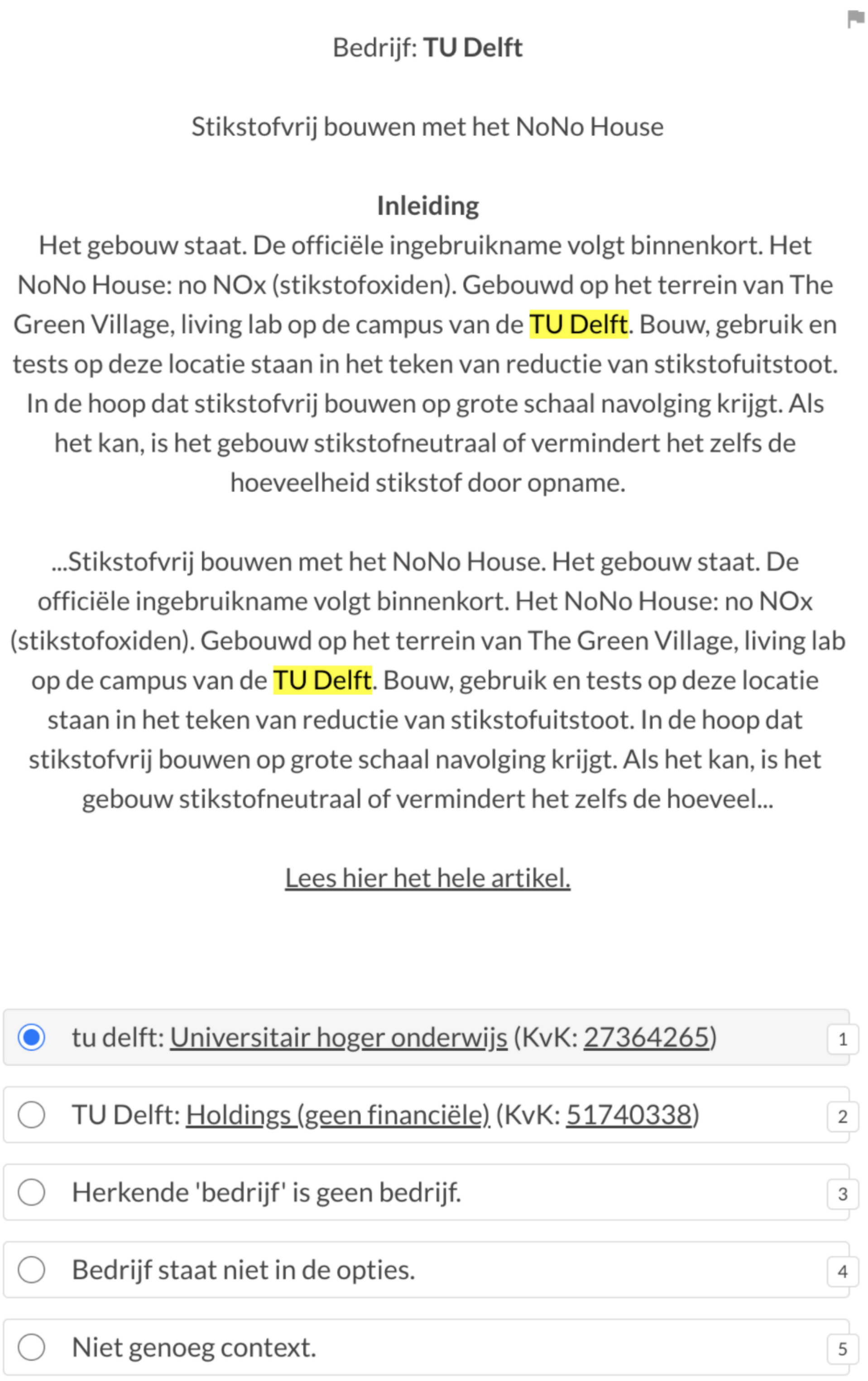
De sectie *Voorbeelden van annotaties* illustreert wanneer elke van de opties gekozen dient te worden.

A.4 Uitzondering: Financiële holdings

Wanneer geen van de SBI-code omschrijvingen van de kandidaten lijkt te passen bij het bedrijf dat genoemd werd in het nieuwsartikel, maar er is wel een kandidaat die qua naam overeenkomt en als SBI-code omschrijving 'Financiële holdings' heeft, dan mag deze optie worden gekozen. Deze kandidaat komt dan dicht in de buurt genoeg. Dit geldt alleen voor de SBI-code omschrijving 'Financiële holdings'.

A.5 Voorbeelden van annotaties

Ter illustratie en verduidelijking van de annotatietaak worden hier een aantal voorbeelden van gewenste annotaties weergegeven.



Bedrijf: **TU Delft**

Stikstofvrij bouwen met het NoNo House

Inleiding

Het gebouw staat. De officiële ingebruikname volgt binnenkort. Het NoNo House: no NOx (stikstofoxiden). Gebouwd op het terrein van The Green Village, living lab op de campus van de TU Delft. Bouw, gebruik en tests op deze locatie staan in het teken van reductie van stikstofuitstoot. In de hoop dat stikstofvrij bouwen op grote schaal navolging krijgt. Als het kan, is het gebouw stikstofneutraal of vermindert het zelfs de hoeveelheid stikstof door opname.

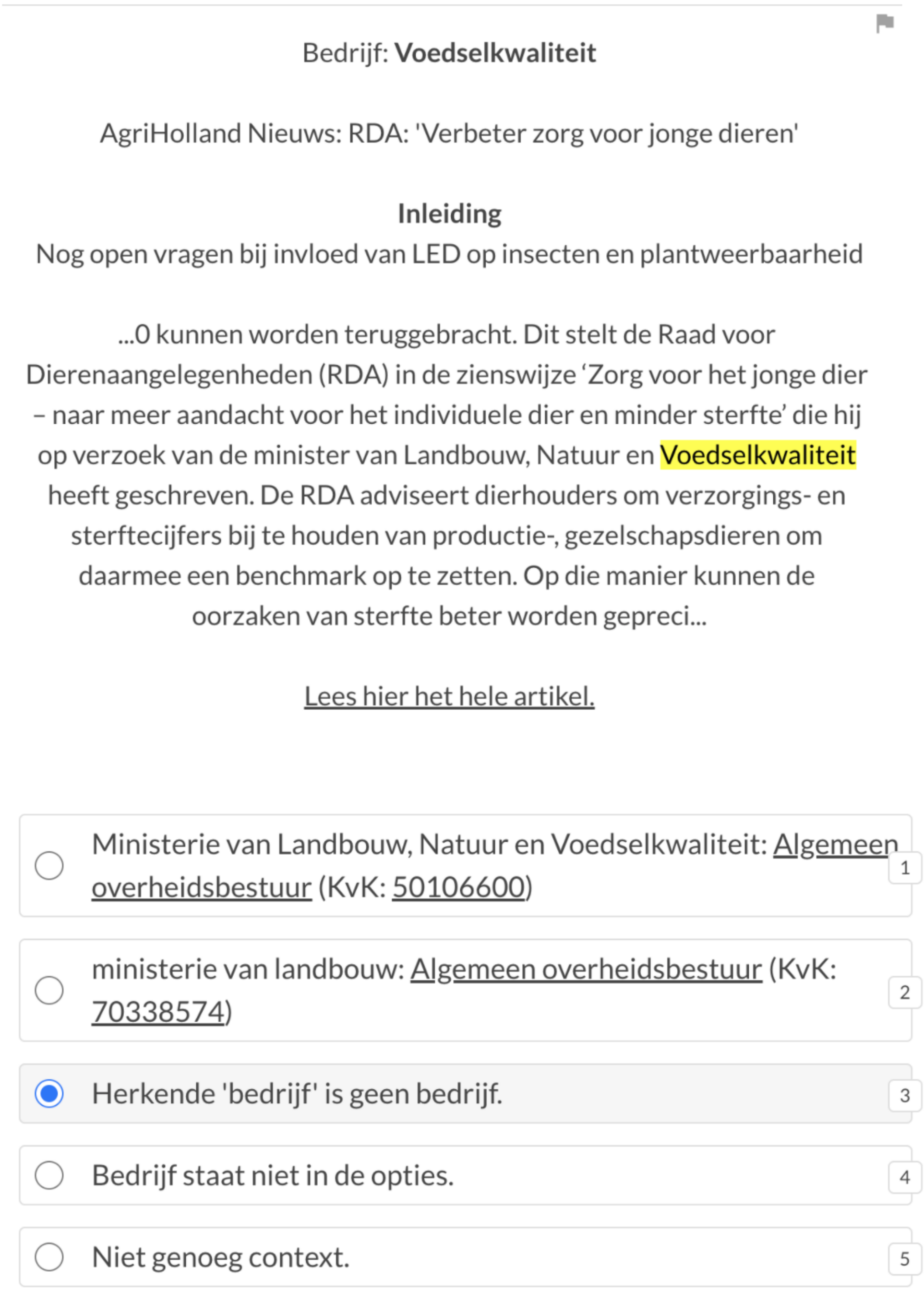
...Stikstofvrij bouwen met het NoNo House. Het gebouw staat. De officiële ingebruikname volgt binnenkort. Het NoNo House: no NOx (stikstofoxiden). Gebouwd op het terrein van The Green Village, living lab op de campus van de TU Delft. Bouw, gebruik en tests op deze locatie staan in het teken van reductie van stikstofuitstoot. In de hoop dat stikstofvrij bouwen op grote schaal navolging krijgt. Als het kan, is het gebouw stikstofneutraal of vermindert het zelfs de hoeveel...

Lees hier het hele artikel.

- ● tu delft: Universitair hoger onderwijs (KvK: 27364265) 1
- ○ TU Delft: Holdings (geen financiële) (KvK: 51740338) 2
- ○ Herkende 'bedrijf' is geen bedrijf. 3
- ○ Bedrijf staat niet in de opties. 4
- ○ Niet genoeg context. 5

Figure A.1: Een annotatie waarbij 1 van de kandidaten de juiste optie is.

In voorbeeld 1 komt de SBI-code omschrijving van de eerste kandidaat het beste overeen met het bedrijf dat genoemd wordt in het artikel. Het woord ‘campus’ geeft aan dat het om de fysieke universiteit van Delft gaat, en de omschrijving ‘Universitair hoger onderwijs’ past daar beter bij dan ‘Holdings (geen financiële)’.



Bedrijf: **Voedselkwaliteit**

AgriHolland Nieuws: RDA: 'Verbeter zorg voor jonge dieren'

Inleiding
Nog open vragen bij invloed van LED op insecten en plantweerbaarheid

...0 kunnen worden teruggebracht. Dit stelt de Raad voor Dierenaangelegenheden (RDA) in de zienswijze 'Zorg voor het jonge dier – naar meer aandacht voor het individuele dier en minder sterfte' die hij op verzoek van de minister van Landbouw, Natuur en Voedselkwaliteit heeft geschreven. De RDA adviseert dierhouders om verzorgings- en sterftecijfers bij te houden van productie-, gezelschapsdieren om daarmee een benchmark op te zetten. Op die manier kunnen de oorzaken van sterfte beter worden gepreci...

Lees hier het hele artikel.

○ Ministerie van Landbouw, Natuur en Voedselkwaliteit: Algemeen overheidsbestuur (KvK: 50106600) 1

○ ministerie van landbouw: Algemeen overheidsbestuur (KvK: 70338574) 2

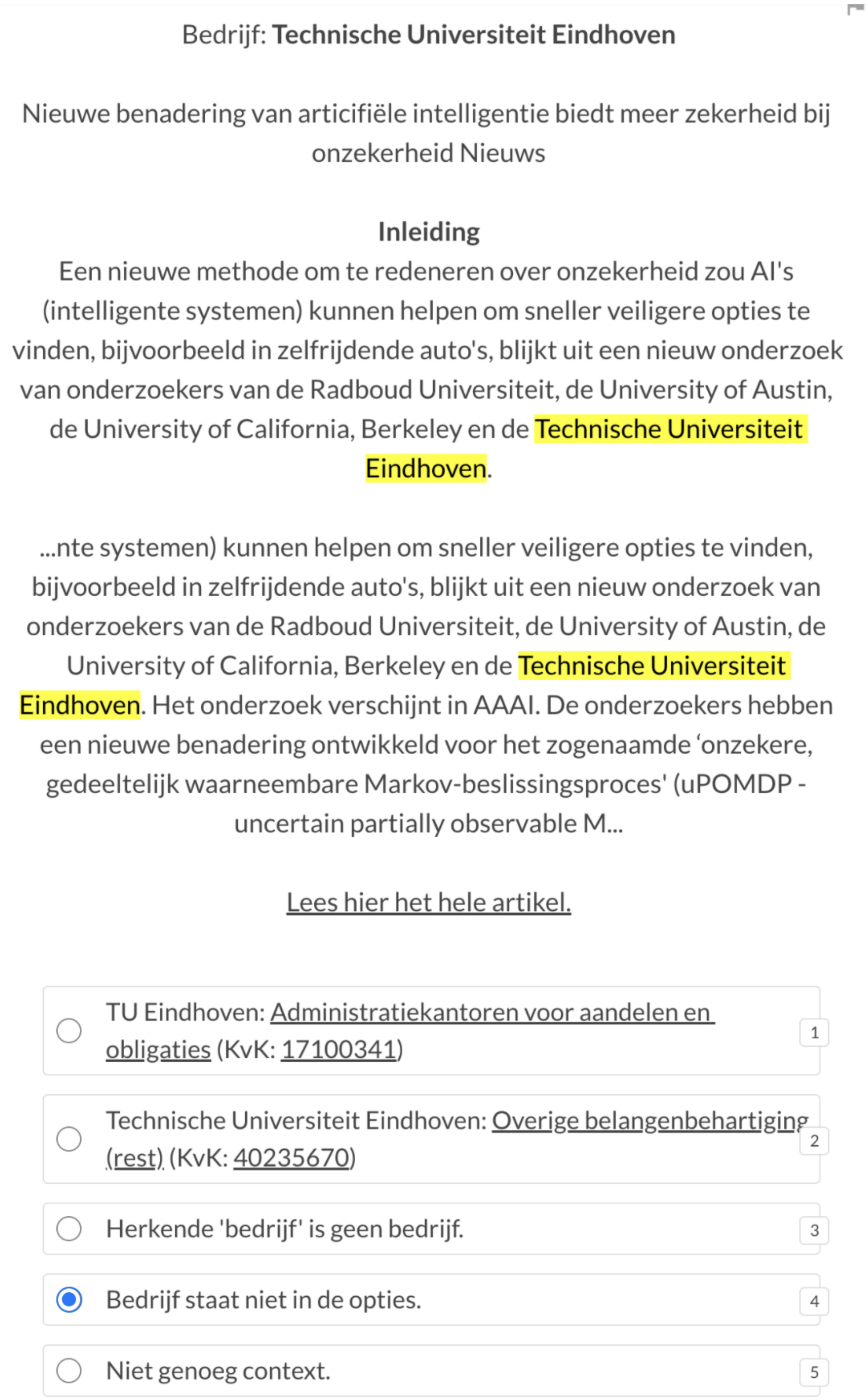
● Herkende 'bedrijf' is geen bedrijf. 3

○ Bedrijf staat niet in de opties. 4

○ Niet genoeg context. 5

Figure A.2: Een annotatie waarbij het bedrijf niet goed herkend is.

In dit voorbeeld wordt ‘Voedselkwaliteit’ incorrect gezien als bedrijfsnaam. De kandidaat die wat betreft naam het meest in de buurt komt, is ‘het ministerie van Landbouw, Natuur en Voedselkwaliteit.’, maar omdat ‘Voedselkwaliteit’ op zichzelf geen bedrijf is, is de juiste keuze hier “Herkende ‘bedrijf’ is geen bedrijf.” Alleen wanneer een volledige bedrijfsnaam herkend wordt in een nieuwsartikel moet hiervoor een kandidaat worden geselecteerd. Ook als er extra woorden bij het herkende bedrijf zitten, zoals “Naast Itho Deelderop” in plaats van “Itho Deelderop” moet de optie “Herkende bedrijf is geen bedrijf.” worden gekozen. Uitzonderingen hierop zijn lidwoorden; wanneer bijvoorbeeld “de Belastingdienst” herkend wordt als bedrijf, maar de naam van de kandidaat is “Belastingdienst”, dan wordt dit gezien als hetzelfde. Andersom geldt dit ook: wanneer ‘de’ deel is van een bedrijfsnaam, maar in het artikel wordt de bedrijfsnaam zonder ‘de’ herkend, dan wordt dit ook gezien als hetzelfde.



Bedrijf: **Technische Universiteit Eindhoven**

Nieuwe benadering van artificiële intelligentie biedt meer zekerheid bij onzekerheid Nieuws

Inleiding

Een nieuwe methode om te redeneren over onzekerheid zou AI's (intelligente systemen) kunnen helpen om sneller veiligere opties te vinden, bijvoorbeeld in zelfrijdende auto's, blijkt uit een nieuw onderzoek van onderzoekers van de Radboud Universiteit, de University of Austin, de University of California, Berkeley en de Technische Universiteit Eindhoven.

...nte systemen) kunnen helpen om sneller veiligere opties te vinden, bijvoorbeeld in zelfrijdende auto's, blijkt uit een nieuw onderzoek van onderzoekers van de Radboud Universiteit, de University of Austin, de University of California, Berkeley en de Technische Universiteit Eindhoven. Het onderzoek verschijnt in AAAI. De onderzoekers hebben een nieuwe benadering ontwikkeld voor het zogenaamde 'onzekere, gedeeltelijk waarneembare Markov-beslissingsproces' (uPOMDP - uncertain partially observable M...

Lees hier het hele artikel.

○ TU Eindhoven: Administratiekantoren voor aandelen en obligaties (KvK: 17100341) 1

○ Technische Universiteit Eindhoven: Overige belangenbehartiging (rest) (KvK: 40235670) 2

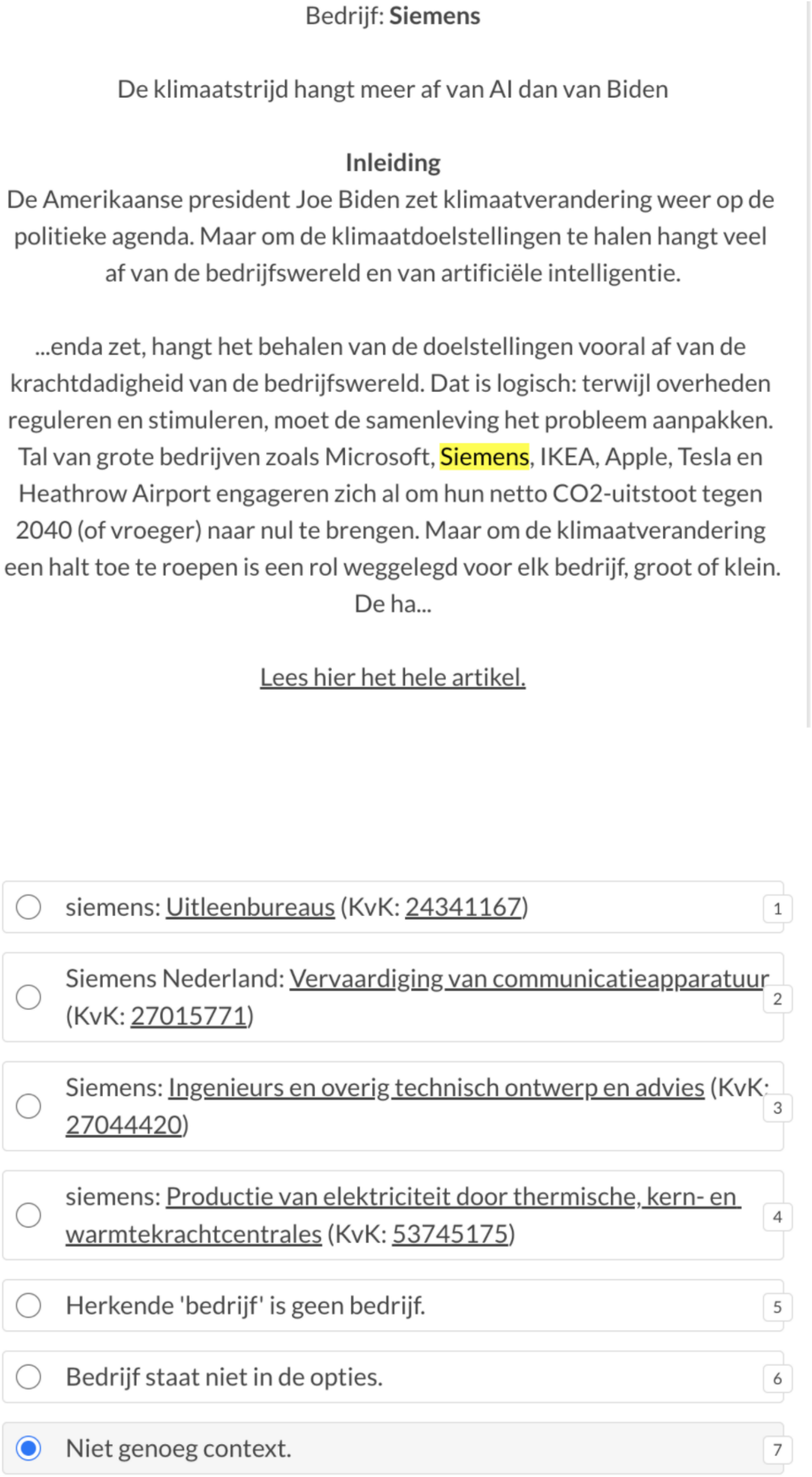
○ Herkende 'bedrijf' is geen bedrijf. 3

● Bedrijf staat niet in de opties. 4

○ Niet genoeg context. 5

Figure A.3: Een annotatie waarbij geen van de kandidaten de juiste kandidaat is.

In dit voorbeeld wordt Technische Universiteit Eindhoven herkend als bedrijf en uit de context is duidelijk dat de universiteit zelf bedoeld wordt. Er zijn 2 kandidaten met dezelfde naam, maar beide hebben SBI-code omschrijvingen die niet overeenkomen met wat het een universiteit is, namelijk een universitaire onderwijsinstelling. Uit het handelsregister blijkt dat er wel een bedrijf bestaat met als naam "Technische Universiteit Eindhoven" en een omschrijving waarin "universitair hoger onderwijs" voorkomt. Als deze tussen de kandidaten stond zou dit de juiste zijn, maar dit is niet het geval, en daarom is "Bedrijf staat niet in de opties" de juiste optie.



Bedrijf: **Siemens**

De klimaatstrijd hangt meer af van AI dan van Biden

Inleiding

De Amerikaanse president Joe Biden zet klimaatverandering weer op de politieke agenda. Maar om de klimaatdoelstellingen te halen hangt veel af van de bedrijfswereld en van artificiële intelligentie.

...enda zet, hangt het behalen van de doelstellingen vooral af van de krachtdadigheid van de bedrijfswereld. Dat is logisch: terwijl overheden reguleren en stimuleren, moet de samenleving het probleem aanpakken. Tal van grote bedrijven zoals Microsoft, Siemens, IKEA, Apple, Tesla en Heathrow Airport engageren zich al om hun netto CO2-uitstoot tegen 2040 (of vroeger) naar nul te brengen. Maar om de klimaatverandering een halt toe te roepen is een rol weggelegd voor elk bedrijf, groot of klein. De ha...

Lees hier het hele artikel.

○ siemens: Uitleenbureaus (KvK: 24341167) 1

○ Siemens Nederland: Vervaardiging van communicatieapparatuur (KvK: 27015771) 2

○ Siemens: Ingenieurs en overig technisch ontwerp en advies (KvK: 27044420) 3

○ siemens: Productie van elektriciteit door thermische, kern- en warmtekrachtcentrales (KvK: 53745175) 4

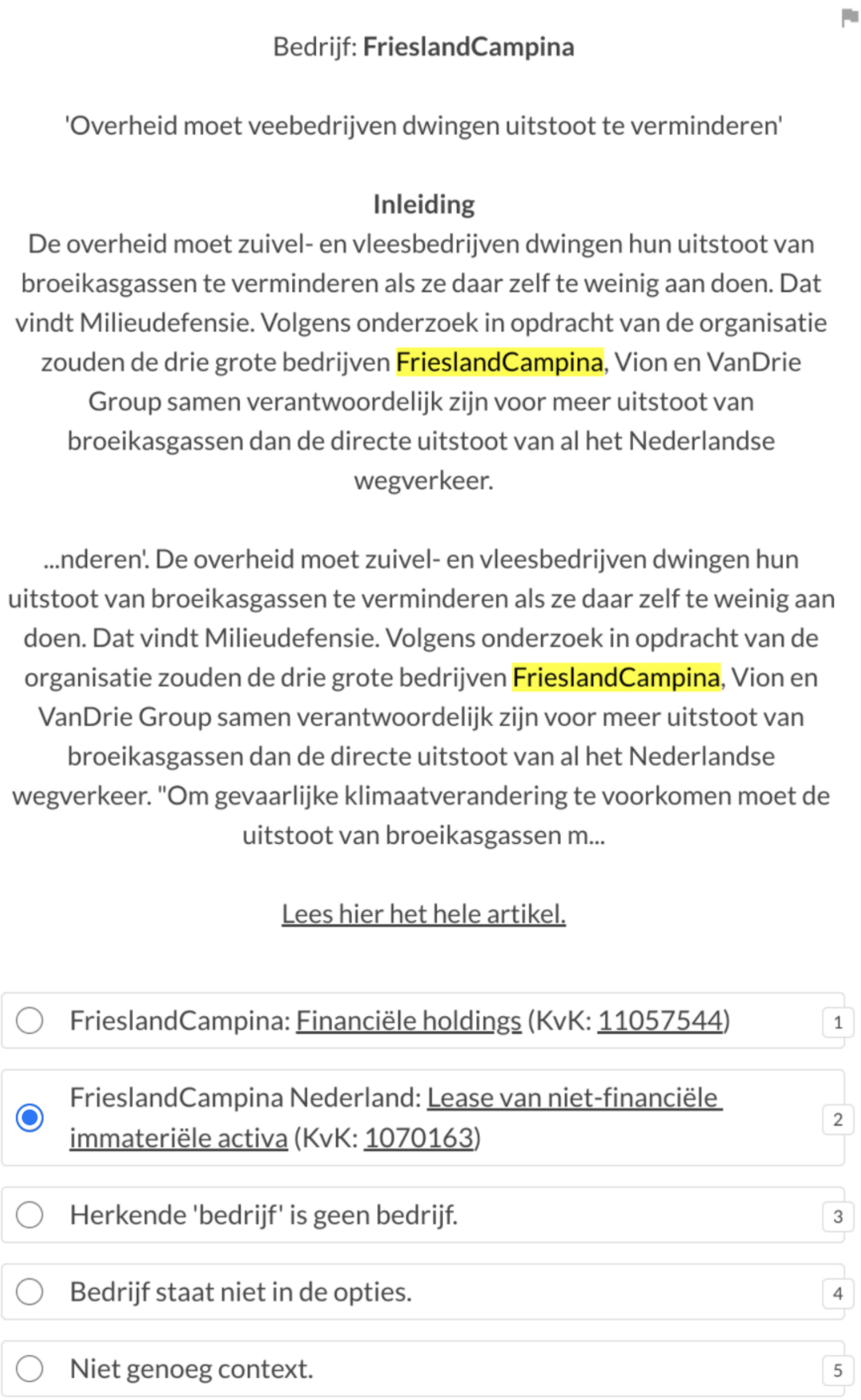
○ Herkende 'bedrijf' is geen bedrijf. 5

○ Bedrijf staat niet in de opties. 6

● Niet genoeg context. 7

Figure A.4: Een annotatie waarbij het nieuwsartikel niet genoeg context biedt om de juiste kandidaat te kiezen.

Voor bedrijfsnaam ‘Siemens’ zijn er vier kandidaten met redelijk verschillende SBI-code omschrijvingen. Siemens wordt genoemd in een opsomming met andere grote bedrijven, die in verschillende gebieden actief zijn en de context van het artikel kan op geen manier duidelijk maken welke kandidaat wordt bedoeld, ook niet als het hele artikel wordt gelezen door op ’Lees hier het hele artikel’. Daarom is de optie ‘Niet genoeg context.’ in dit geval de juiste keuze.



Bedrijf: **FrieslandCampina**

'Overheid moet veebedrijven dwingen uitstoot te verminderen'

Inleiding

De overheid moet zuivel- en vleesbedrijven dwingen hun uitstoot van broeikasgassen te verminderen als ze daar zelf te weinig aan doen. Dat vindt Milieudefensie. Volgens onderzoek in opdracht van de organisatie zouden de drie grote bedrijven FrieslandCampina, Vion en VanDrie Group samen verantwoordelijk zijn voor meer uitstoot van broeikasgassen dan de directe uitstoot van al het Nederlandse wegverkeer.

...nderen'. De overheid moet zuivel- en vleesbedrijven dwingen hun uitstoot van broeikasgassen te verminderen als ze daar zelf te weinig aan doen. Dat vindt Milieudefensie. Volgens onderzoek in opdracht van de organisatie zouden de drie grote bedrijven FrieslandCampina, Vion en VanDrie Group samen verantwoordelijk zijn voor meer uitstoot van broeikasgassen dan de directe uitstoot van al het Nederlandse wegverkeer. "Om gevaarlijke klimaatverandering te voorkomen moet de uitstoot van broeikasgassen m...

Lees hier het hele artikel.

○ FrieslandCampina: Financiële holdings (KvK: 11057544) 1

● FrieslandCampina Nederland: Lease van niet-financiële immateriële activa (KvK: 1070163) 2

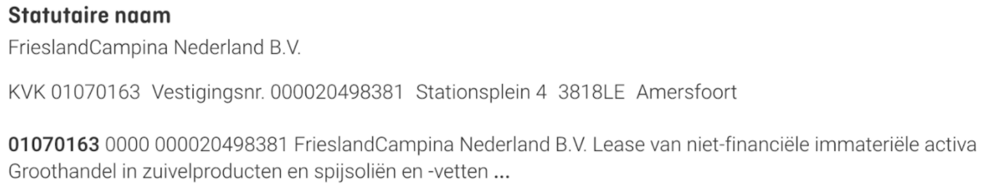
○ Herkende 'bedrijf' is geen bedrijf. 3

○ Bedrijf staat niet in de opties. 4

○ Niet genoeg context. 5

Figure A.5: Een annotatie waarbij de juiste optie niet direct duidelijk is.

FrieslandCampina is een redelijk bekend bedrijf dat gespecialiseerd is in zuivel en de kans is groot dat een annotator hiermee bekend is. Op basis van de SBI-code omschrijvingen lijkt geen van deze kandidaten de juiste maar als het wordt opgezocht in het Handelsregister, is het eerste resultaat als volgt:



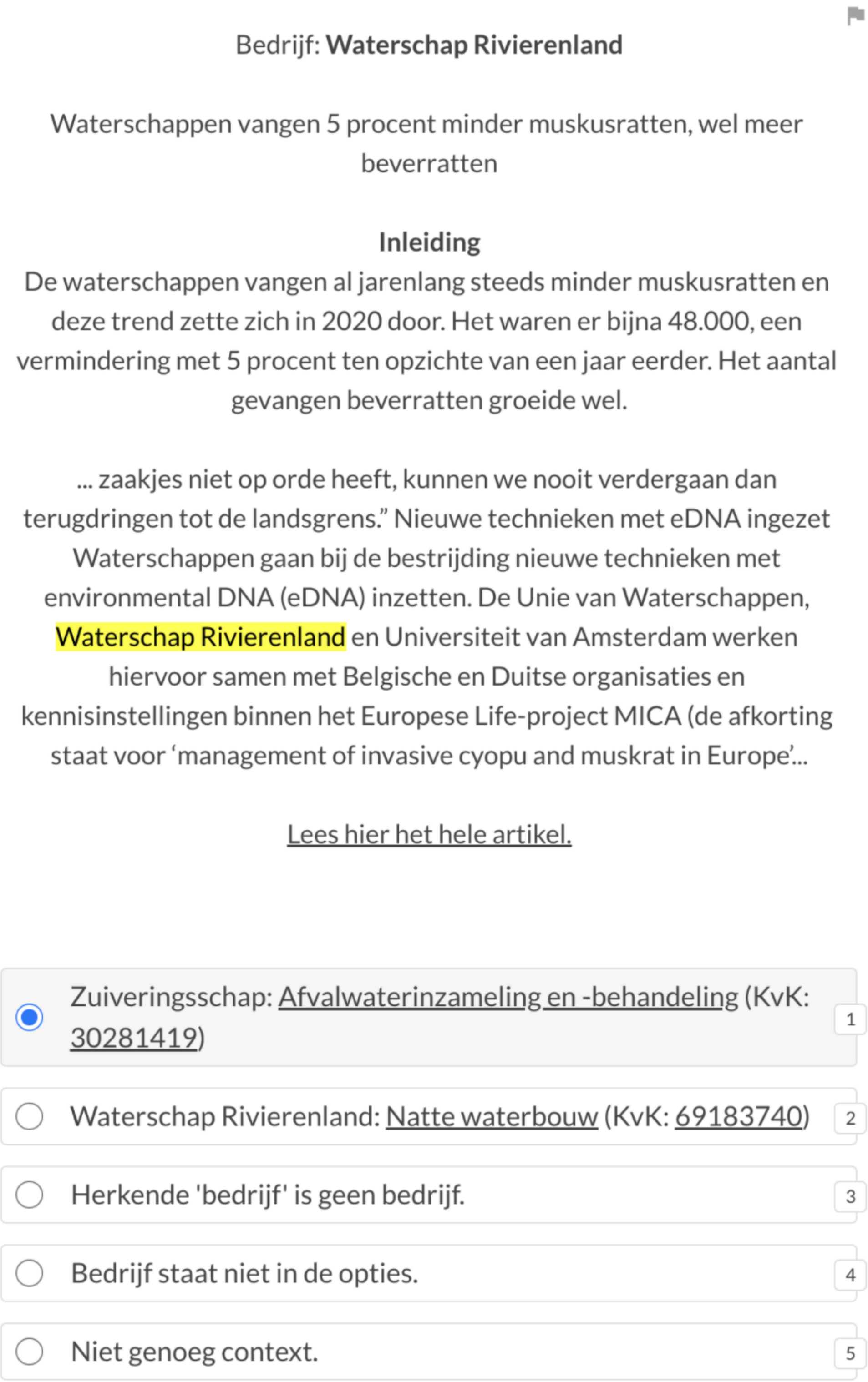
Statutaire naam
FrieslandCampina Nederland B.V.

KVK 01070163 Vestigingsnr. 000020498381 Stationsplein 4 3818LE Amersfoort

01070163 0000 000020498381 FrieslandCampina Nederland B.V. Lease van niet-financiële immateriële activa Groothandel in zuivelproducten en spijsoliën en -vetten **...**

Figure A.6: Resultaat van het KvK-nummer van de tweede kandidaat in het Handelsregister.

De SBI-code omschrijving van de kandidaat in Prodigy komt overeen met het resultaat van het Handelsregister, maar er staat nog een omschrijving: “Groothandel in zuivelproducten en spijsoliën en vetten.” Deze omschrijving lijkt beter te passen bij FrieslandCampina en het KvK-nummer komt overeen met kandidaat 2, wat aangeeft dat kandidaat 2 in dit geval de juiste keuze is.



Bedrijf: **Waterschap Rivierenland**

Waterschappen vangen 5 procent minder muskusratten, wel meer beverratten

Inleiding

De waterschappen vangen al jarenlang steeds minder muskusratten en deze trend zette zich in 2020 door. Het waren er bijna 48.000, een vermindering met 5 procent ten opzichte van een jaar eerder. Het aantal gevangen beverratten groeide wel.

... zaakjes niet op orde heeft, kunnen we nooit verdergaan dan terugdringen tot de landsgrens." Nieuwe technieken met eDNA ingezet Waterschappen gaan bij de bestrijding nieuwe technieken met environmental DNA (eDNA) inzetten. De Unie van Waterschappen, Waterschap Rivierenland en Universiteit van Amsterdam werken hiervoor samen met Belgische en Duitse organisaties en kennisinstellingen binnen het Europese Life-project MICA (de afkorting staat voor 'management of invasive cyopu and muskrat in Europe'...

Lees hier het hele artikel.

● Zuiveringsschap: Afvalwaterinzameling en -behandeling (KvK: 30281419) 1

○ Waterschap Rivierenland: Natte waterbouw (KvK: 69183740) 2

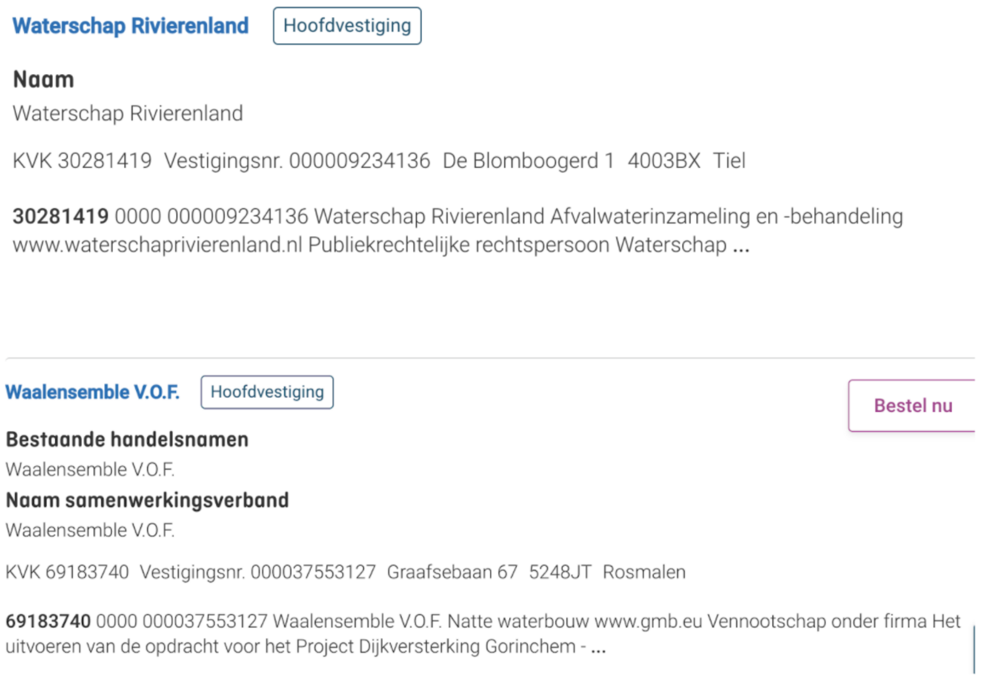
○ Herkende 'bedrijf' is geen bedrijf. 3

○ Bedrijf staat niet in de opties. 4

○ Niet genoeg context. 5

Figure A.7: Een annotatie waarbij de naam van de juiste kandidaat niet direct overeen lijkt te komen met het bedrijf dat genoemd werd.

Het laatste voorbeeld illustreert dat de basisnamen van de kandidaten vaak niet overeenkomen met de basisnaam in het Handelsregister van de KvK. In dit artikel wordt het bedrijf "Waterschap Rivierenland" herkend. Er is een kandidaat die precies deze naam heeft, en de SBI-code omschrijving lijkt overeen te komen met de context van het artikel, wat doet vermoeden dat dit de juiste kandidaat is. Wanneer de KvK-nummers worden ingevoerd in het zoeksysteem van het Handelsregister, worden respectievelijk de volgende resultaten getoond (vergelijk de KvK-nummers):



Waterschap Rivierenland

Naam

Waterschap Rivierenland

KVK 30281419 Vestigingsnr. 000009234136 De Blomboogerd 1 4003BX Tiel

30281419 0000 000009234136 Waterschap Rivierenland Afvalwaterinzameling en -behandeling www.waterschaprivierenland.nl Publiekrechtelijke rechtspersoon Waterschap ...

Waalensemble V.O.F.

Bestaande handelsnamen

Waalensemble V.O.F.

Naam samenwerkingsverband

Waalensemble V.O.F.

KVK 69183740 Vestigingsnr. 000037553127 Graafsebaan 67 5248JT Rosmalen

69183740 0000 000037553127 Waalensemble V.O.F. Natte waterbouw www.gmb.eu Vennootschap onder firma Het uitvoeren van de opdracht voor het Project Dijkversterking Gorinchem - ...

Figure A.8: Resultaten in het Handelsregister voor beide kandidaten voor de genoemde bedrijfsnaam.

Omdat de informatie uit het Handelsregister leidend is, is de kandidaat met KvK-nummer 30281419 in dit geval de juiste kandidaat. Gevallen zoals deze worden veroorzaakt door fouten in de database met bedrijfsinformatie die wordt gebruikt voor deze database en zijn moeilijk te herkennen.

Bibliography

Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., & Hellmann, S. (2009). Dbpedia-a crystallization point for the web of data. *Journal of web semantics*, *7*(3), 154–165.

Broscheit, S. (2020). Investigating entity knowledge in bert with simple neural end-to-end entity linking. *arXiv preprint arXiv:2003.05473*.

Bryl, V., Bizer, C., & Paulheim, H. (2015). Gathering alternative surface forms for dbpedia entities. In *NLP-DBPEDIA@ ISWC*, (pp. 13–24).

Dalton, J., & Dietz, L. (2013). A neighborhood relevance model for entity linking. In *Proceedings of the 10th Conference on Open Research Areas in Information Retrieval*, (pp. 149–156). Citeseer.

de Vries, W., van Cranenburgh, A., Bisazza, A., Caselli, T., van Noord, G., & Nissim, M. (2019). Bertje: A dutch bert model. *arXiv preprint arXiv:1912.09582*.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Durrett, G., & Klein, D. (2014). A joint model for entity analysis: Coreference, typing, and linking. *Transactions of the association for computational linguistics*, *2*, 477–490.

Finkel, J. R., Grenager, T., & Manning, C. D. (2005). Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, (pp. 363–370).

Ganea, O.-E., & Hofmann, T. (2017). Deep joint entity disambiguation with local neural attention. *arXiv preprint arXiv:1704.04920*.

Giles, C. L., Zha, H., & Han, H. (2005). Name disambiguation in author citations using a k-way spectral clustering method. In *Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries (JCDL'05)*, (pp. 334–343). IEEE.

Gregg, F., & Eder, D. (2019). Project title. https://github.com/dedupeio/dedupe.

Guo, S., Chang, M.-W., & Kiciman, E. (2013). To link or not to link? a study on end-to-end tweet entity linking. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, (pp. 1020–1030).

Hachey, B., Radford, W., Nothman, J., Honnibal, M., & Curran, J. R. (2013). Evaluating entity linking with wikipedia. *Artificial Intelligence*, *194*, 130–150. Artificial Intelligence, Wikipedia and Semi-Structured Resources.
URL https://www.sciencedirect.com/science/article/pii/S0004370212000446

Han, H., Zha, H., & Giles, C. L. (2005). Name disambiguation in author citations using a k-way spectral clustering method. *Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL '05)*, (pp. 334–343).

Hasibi, F., Balog, K., & Bratsberg, S. E. (2016). Exploiting entity linking in queries for entity retrieval. In *Proceedings of the 2016 acm international conference on the theory of information retrieval*, (pp. 209–218).

Hendriks, B., Groth, P., & van Erp, M. (2021). Recognising and linking entities in old dutch texts: A case study on voc notary records.

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 2011 Conference on Empirical Methods in Natural Language Processing*, (pp. 782–792).

Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). spacy: Industrial-strength natural language processing in python.
URL https://doi.org/10.5281/zenodo.1212303

Kolitsas, N., Ganea, O.-E., & Hofmann, T. (2018). End-to-end neural entity linking. *arXiv preprint arXiv:1808.07699*.

Martins, P. H., Marinho, Z., & Martins, A. F. (2019). Joint learning of named entity recognition and entity linking. *arXiv preprint arXiv:1907.08243*.

McNamee, P., & Dang, H. T. (2009). Overview of the tac 2009 knowledge base population track. In *Text Analysis Conference (TAC)*, vol. 17, (pp. 111–113).

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*, (pp. 1–8).

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*, (pp. 233–242).

Miller, G. A. (1995). Wordnet: a lexical database for english. *Communications of the ACM*, *38*(11), 39–41.

Mulang', I. O., Singh, K., Prabhu, C., Nadgeri, A., Hoffart, J., & Lehmann, J. (2020). Evaluating the impact of knowledge graph context on entity disambiguation models. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, (pp. 2157–2160).

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

Navigli, R. (2009). Word sense disambiguation: A survey. *ACM computing surveys (CSUR)*, *41*(2), 1–69.

Nguyen, D. B., Theobald, M., & Weikum, G. (2016). J-nerd: joint named entity recognition and disambiguation with rich linguistic features. *Transactions of the Association for Computational Linguistics*, *4*, 215–229.

Olieman, A., Azarbonyad, H., Dehghani, M., Kamps, J., & Marx, M. (2014). Entity linking by focusing dbpedia candidate entities. In *Proceedings of the first international workshop on Entity recognition & disambiguation*, (pp. 13–24).

Qureshi, M. A., O'Riordan, C., & Pasi, G. (2014). Exploiting wikipedia for entity name disambiguation in tweets. In *International Conference on Applications of Natural Language to Data Bases/Information Systems*, (pp. 184–195). Springer.

Rao, D., McNamee, P., & Dredze, M. (2013). Entity linking: Finding extracted entities in a knowledge base. In *Multi-source, multilingual information extraction and summarization*, (pp. 93–115). Springer.

Röder, M., Usbeck, R., & Ngonga Ngomo, A.-C. (2018). Gerbil–benchmarking named entity recognition and linking consistently. *Semantic Web*, *9*(5), 605–625.

Shen, W., Wang, J., & Han, J. (2014). Entity linking with a knowledge base: Issues, techniques, and solutions. *IEEE Transactions on Knowledge and Data Engineering*, *27*(2), 443–460.

Sil, A., & Yates, A. (2013). Re-ranking for joint named-entity recognition and linking. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, (pp. 2369–2374).

Smith, S. L., Kindermans, P.-J., Ying, C., & Le, Q. V. (2017). Don't decay the learning rate, increase the batch size. *arXiv preprint arXiv:1711.00489*.

Spina, D., Amigó, E., & Gonzalo, J. (2011). Filter keywords and majority class strategies for company name disambiguation in twitter. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, (pp. 50–61). Springer.

Spina, D., Gonzalo, J., & Amigó, E. (2013). Discovering filter keywords for company name disambiguation in twitter. *Expert Systems with Applications*, *40*(12), 4986–5003.

Tsai, C.-T., & Roth, D. (2016). Cross-lingual wikification using multilingual embeddings. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, (pp. 589–598).

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*, (pp. 1133–1143).

Van Noord, G., Bouma, G., Van Eynde, F., De Kok, D., Van der Linde, J., Schuurman, I., Sang, E. T. K., & Vandeghinste, V. (2013). Large scale syntactic annotation of written dutch: Lassy. In *Essential speech and language technology for Dutch*, (pp. 147–164). Springer, Berlin, Heidelberg.

van Veen, T., Lonij, J., & Faber, W. (2016). Linking named entities in dutch historical newspapers. In *Research Conference on Metadata and Semantics Research*, (pp. 205–210). Springer.

Zhang, S., Wu, J., Zheng, D., Meng, Y., & Yu, H. (2012). An adaptive method for organization name disambiguation with feature reinforcing. In *Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation*, (pp. 237–245).