Consolidating Heterogeneous Enterprise Data for Named Entity Linking and Web Intelligence

Albert Weichselbraun[1], Daniel Streiff[1], and Arno Scharl[2]

[1]Swiss Institute for Information Research, University of Applied Sciences, Pulvermühlestrasse 57, CH-7004 Chur, Switzerland, {albert.weichselbraun,daniel.streiff}@htwchur.ch
[2]Department of New Media Technology, MODUL University Vienna, Am Kahlenberg 1, A-1190 Vienna, Austria, scharl@modul.ac.at

April 28, 2015

Abstract

Linking named entities to structured knowledge sources paves the way for state-of-the-art Web intelligence applications which assign sentiment to the correct entities, identify trends, and reveal relations between organizations, persons and products. For this purpose this paper introduces Recognyze, a named entity linking component that uses background knowledge obtained from linked data repositories, and outlines the process of transforming heterogeneous data silos within an organization into a linked enterprise data repository which draws upon popular linked open data vocabularies to foster interoperability with public data sets. The presented examples use comprehensive real-world data sets from Orell Füssli Business Information, Switzerland's largest business information provider. The linked data repository created from these data sets comprises more than nine million triples on companies, the companies' contact information, key people, products and brands. We identify the major challenges of tapping into such sources for named entity linking, and describe required data pre-processing techniques to use and integrate such data sets, with a special focus on disambiguation and ranking algorithms. Finally, we conduct a comprehensive evaluation based on business news from the *New Journal of Zurich* and *AWP Financial News* to illustrate how these techniques improve the performance of the Recognyze named entity linking component.

Keywords: *linked open data; linked enterprise data; named entity linking; named entity resolution; business news; Web intelligence; data pre-processing; data consolidation*

1 Introduction

The term *business intelligence* gained prominence in the 1990s when Howard Dresner from the Gartner Group started using it in its current interpretation[28]. Business intelligence is an umbrella term that describes concepts and methods to improve decision making through fact-based support systems[4] - combining data acquisition, data storage and knowledge management components with analytical methods for processing large amounts of data and providing decision makers with timely and high-quality input to support their decision processes[20].

User-generated content from social media platforms has become a valuable source of feedback that sheds light on a company's business operations, helps to optimize communication strategies and marketing campaigns, and supports the customization of products and services to consumer needs. The significant potential of user-generated content motivated companies to apply automated content analysis to blogs, product reviews and social media streams. State-of-the-art Web intelligence systems use data mining engines to extract structured knowledge from such unstructured textual sources[3]. Named entity linking is a crucial task in this process, ensuring that the extracted items are assigned to the correct entities (people, organizations, products, etc.).

Preprint of an article submitted to the International Journal on Artificial Intelligence Tools

The publisher's version of the article is available at http://dx.doi.org/10.1142/S0218213015400084.

This paper introduces Recognyze as a novel component for this linking process, drawing upon background knowledge from structured sources - e.g., linked data repositories such as DPpedia, Freebase and GeoNames. As part of the webLyzard Web intelligence platform (www.weblyzard-.com), Recognyze has been adopted to identify and link persons, organizations and locations in two environmental Web applications: (i) the Media Watch on Climate Change shown in Figure 1 (www.ecoresearch.net/climate) - a multilingual visual analytics portal to explore climate change coverage from English, French and German online sources, and (ii) the Climate Resilience Toolkit (toolkit.climate.gov) - a new decision support initiative by the U.S. Government that provides expert knowledge to help citizens and communities manage climate-related risks and opportunities.

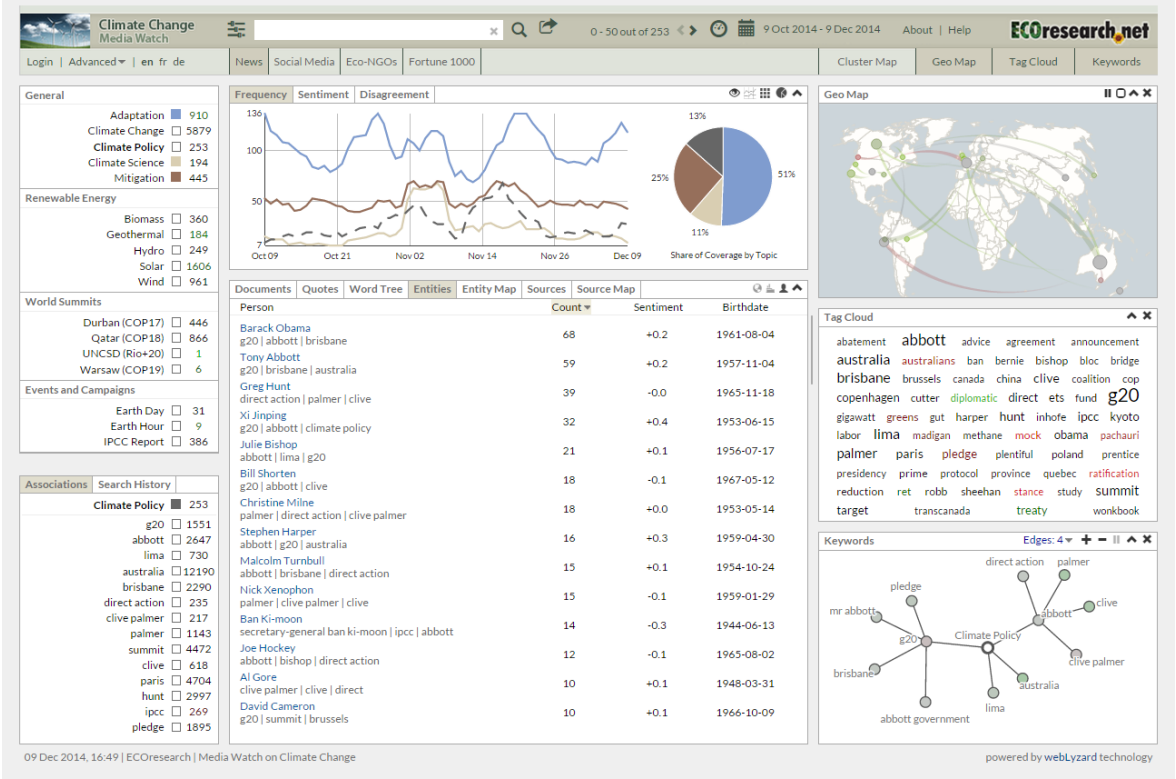

Figure 1: Screenshot of the Media Watch on Climate Change (www.ecoresearch.net/climate) showing a list of people associated with "energy policy" based on Anglo-American news media coverage between November and December 2014

As demonstrated by these two use cases, named entity linking unfolds its full potential once the identified entities are linked to existing knowledge repositories such as the above mentioned linked data sources, or a company's proprietary business directories. In a business context, however, information is often distributed across heterogeneous systems which act as isolated data silos with little interaction among each other. Omitola et al.[22] observed a number of challenges when transforming such silos into interoperable linked data sets: (i) establishing a business case, (ii) data discovery and provenance, (iii) information extraction, (iv) data cleaning, (v) data interlinking, (vi) modelling and (vii) data management. Addressing these challenges has clear economic advantages. Wolters Kluwer, one of the world's largest legal publishers, for instance, has build a standard representation of legal content which allows an integration of publishable assets across the company[16]. DataPatrol[1] is an early warning system which aims at protecting its customers from identity theft. The company monitors publicly available and privately acquired data for indicators of online identity theft and financial fraud, alerting users once critical information has been spotted and suggesting countermeasures for protecting themselves. DataPatrol draws upon linked enterprise data for customer service, provenance, and debugging purposes, but also publishes linked open data on a person's digital status[12].

Interlinking company data with public data sets blends these data sets with semantic background knowledge, encoded based on well-defined vocabularies and maintained by third parties. A deeper

[1] http://www.garlik.com/datapatrol

integration of named entity linking and business data sets allows automatically linking intelligence obtained from user-generated content to the corresponding business data sets.

This article is structured as follows: Section 2 discusses related work on named entity linking and information extraction approaches that leverage background knowledge for increased accuracy. Section 3 then describes the process of identifying relevant databases within a company and transforming them into interoperable linked enterprise data. Section 4 outlines the characteristics and challenges of using these data for named entity linking, including the specific pre-processing and disambiguation techniques of Recognyze. A detailed evaluation in Section 5 is followed by an outlook and conclusions in Section 6.

2 Related Work

2.1 Named entity recognition

Named entity recognition (NER) identifies references to named entities in unstructured documents and classifies them into categories such as people, organizations and locations. Urbansky et al.[26] distinguish between three approaches towards named entity recognition: (i) the use of hand-crafted rules or knowledge sources such as lexicons, (ii) supervised machine learning, and (iii) unsupervised machine learning techniques. State of the art named entity recognition tools such as the Stanford NER Tagger[8] distinguish only between a relatively small set of entity categories. HYENA[34], in contrast, draws upon YAGO[14] and sources such as Wikipedia and DBpedia[19] to perform a very fine-grained typing and even considers multi-labeling (e.g. Albert Eistein is a person, a scientist, a theoretical physicists, novel price winner, etc.). Other approaches for fine-grained named entity recognition often use named entity linking and then query the knowledge base to infer the corresponding types.

2.2 Named entity linking

Named entity linking, which is also known as named entity resolution, in contrast, not only classifies named entities but also grounds them to a knowledge base such as DBpedia and Wikipedia, or to a relational database. Gangemi[9] provides an overview of knowledge extraction tools including specific applications for named entity recognition and linking. Wang et al.[27] approach the disambiguation problem by suggesting a graph-based model (MentionRank), which leverages the principle that homogeneous groups of entities often occur in similar documents. When applied to information technology companies, for instance, context-awareness helps to disambiguate terms such as "Apple" or "HP" when they occur in documents with an information technology or business focus.

Many approaches either use Wikipedia for training their models[18, 21] or use background knowledge from Wikipedia to improve the accuracy of the named entity disambiguation process[11, 23, 15]. Han and Zhao[11] observe that leveraging semantic knowledge from Wikipedia yields an improvement of 10.7% over traditional bag-of-word approaches, and a 16.7% improvement over traditional social network-based disambiguation methods.

Pilz and Paaß[23] use a thematic information measure derived from Latent Dirichlet Allocation (LDA) to compare mentions with candidate entities in Wikipedia. Distance metrics in a supervised classification setting enable them to identify the best fitting entity for that particular mention. Kataria et al.[18] use a hierarchical variant of LDA models for named entity disambiguation. They present a semi-supervised hierarchical model that considers Wikipedia to learn name-entity associations, exploits Wikipedia annotations, and uses Wikipedia's category hierarchy for capturing co-occurrence probabilities among entities.

Recently, Nothman et al.[21] used Wikipedia to create multilingual training data for named entity linking tasks, resulting in millions of annotations in nine languages. An evaluation of their Wikipedia-trained models based on English, German, Spanish, Dutch and Russian reference data from the Conference on Natural Language Learning (CONLL) shared task[24, 25] shows that they outperform a number of other approaches to automatic named entity linking.

Fernández et al.[7] present IdentityRank, a supervised algorithm for disambiguating names in news coverage. The authors process historical co-occurrence information on entities and topics, and temporal information on entities prevalent in news streams for estimating the probability of a name to refer to a certain entity. Jung[17] explores how named entity linking methods can be applied to

challenging data sets such as those derived from social media streams, which are characterized by short and often noisy text.

DBpedia Spotlight[5] is another named entity linking system which uses DBpedia to annotate entities in text documents, but lacks advanced pre-processing and is currently limited to DBpedia only. Finally, AIDA[15] is a well-known system for named entity linking which harnesses context information from structured data sources such as DBpedia and YAGO, and introduces a new form of coherence graph that combines the prior probability of an entity being mentioned with context similarity and the coherence among candidate entities for all names referenced in a document. AIDA even supports entities which are not yet covered in the knowledge base (called unlinkable entities, emerging entities or out-of-knowledge-base entities) by introducing an additional unlinkable entity for each mention and computing characteristic phrases for the unlinkable entity prior to disambiguation[13].

2.3 Background knowledge for information extraction

Hoffart et al.[15] and Weichselbraun et al.[31, 32] demonstrate that considering external knowledge for information extraction tasks such as named entity linking can significantly improve the accuracy of the deployed methods.

The field of Natural Language Processing (NLP) has a long history of dealing with the subtleties of human languages. NLP researchers have created comprehensive structured resources that represent common sense knowledge and contain information on ambiguous concepts and potential sentiment indicators. Examples of such resources include ConceptNet[2], SenticNet[3] and SentiWordNet[4]. Recent research in this area shows how methods that have been enhanced with the ability to use background knowledge are able to (i) adapt their evaluations to the text's context[30, 6], (ii) distinguish between ambivalent concepts[31] and, therefore, (iii) provide a much better assessment of the text's sentiment.

Machine learning approaches that limit the use of background knowledge to the training set have also been successful. Wu and Weld[33] use Wikipedia infobox attributes extracted from a cleaned set of infoboxes provided by DBpedia to generate training examples for their information extraction component. They report an improvement of the F-measure of between 18% and 34% when compared to a similar approach that solely relied on hand crafted heuristics for generating training data.

The Recognyze approach presented in this paper uses background knowledge from arbitrary linked data sources to disambiguate and link named entities. In contrast to other methods that only provide basic means to manipulate the data acquired from external knowledge sources, Recognyze offers an advanced infrastructure for validating and enriching these data. Its pre-processing pipeline allows extracting and manipulating names, context information, structural information and an entity's relative importance from the knowledge sources under consideration.

3 Consolidating Heterogeneous Data Repositories

The named entity linking component introduced in this paper uses linked data repositories for disambiguating and linking named entities. This section describes the process of combining and consolidating existing data sets within an organization to create a data repository for named entity linking, and discusses the potential of combining such resources with linked open data repositories such as GeoNames and DBpedia.

3.1 Linked enterprise data

Enterprises often hold their data in heterogeneous and rather isolated data silos that are only accessible through data- and application-specific interfaces. Applying the principles of linked open data to enterprise data is an interesting new research area that promises an integration and consolidation of heterogeneous data sources - e.g., blending private enterprise data with publicly available and *maintained* resources by reusing well-known vocabularies such as Friend-of-a-Friend (FOAF), Dublin Core (DC) and Simple Knowledge Organization System (SKOS).

In a first step, we identified databases and tables with relevant data and obtained the corresponding data dumps in collaboration with Orell Füssli Business Information (OFWI), Switzerland's largest

[2] conceptnet5.media.mit.edu
[3] sentic.net
[4] sentiwordnet.isti.cnr.it

Table 1: Namespaces used for entities extracted from the data dumps.

Entity type	**Namespace URL**
Kompass companies	http://www.semanticlab.net/proj/wisdom/ofwi/kompass/company/
Teledata companies	http://www.semanticlab.net/proj/wisdom/ofwi/teledata/company/
products	http://www.semanticlab.net/proj/wisdom/ofwi/productgroup/
key employees	http://www.semanticlab.net/proj/wisdom/ofwi/person/
company addresses	http://www.semanticlab.net/proj/wisdom/ofwi/address/
industry sectors	http://www.semanticlab.net/proj/wisdom/ofwi/industry/

Table 2: Vocabulary used for the Orell Füssli linked enterprise data repository.

Namespace	**number of elements**	**examples**
dbprop	4	products, distributor, keyPeople, revenue
dbprop-de	1	unternehmensform
dbpedia-owl	5	Company, abstract, industry, numberOfEmployees
foaf	4	Person, firstName, lastName, gender
owl	1	sameAs
schema-org	6	PostalAddress, address, email, faxNumber
ofwi	1	companyStatus

provider of business information. Afterwards, drill-down analyses revealed (i) the following six candidate entity types for extraction from the data dumps: company data from the *Kompass* and *Teledata* databases, data on products and services offered by these companies, the companies' management, addresses, and the industries in which they operate; (ii) database fields which may contain multiple values such as lists of companies or industries; (iii) proprietary codes such as language ids, letters indicating a person's gender, or a company's status; (iv) the used text encodings; and (v) necessary text cleanup transformations.

To ensure interoperability with public resources, we decided to use well-known linked open data vocabularies to describe extracted entities wherever possible and assigned unique namespaces (Table 1) to all extracted entity types.

Table 2 presents an overview of the used vocabularies, the total number of elements taken from a particular vocabulary, and a number of selected example elements. The repository is restricted to vocabulary from public namespaces, with the exception of the `ofwi` namespace that is used to represent a company's status according to OFWI's internal classification schema. For the company's legal form we use the `dbprop-de` rather than `dbprop` namespace because the translation of company types between languages and countries is problematic due to different legal settings.

A python-based information extraction component, translated database fields to RDF triples which describe the corresponding entities, based on mappings specified by domain experts, split multi value fields into multiple atomic statements, and converted proprietary ids to human readable standard mappings. The component also ensured that all text fields are cleaned (i.e. multiple spaces and proprietary control codes removed) and UTF-8 encoded. Figure 2 presents an excerpt of linked data generated for a Swiss optical equipment company, *Carl Zeiss Vision AG*. Company entities from different databases (Kompass and Teledata) have been linked using the `owl:sameAs` predicate (line 12). The process described above yielded background information on additional company names (lines 3-4 and 14), the company's senior managers (lines 34-35), products (lines 28-31), address and contact information (lines 20-25 and 38-43), business figures such as number of employees and turnover (lines 5-6), brands offered by the company (line 29), and the industry sector the company operates in (lines 47-50). Removing duplicates and references to inactive companies yielded a linked enterprise data repository with more than 9 million triples.

The extracted RDF files comprised data on more than 2.9 million companies which have been uploaded to an OpenRDF Sesame server[5].

[5] www.openrdf.org

```
# data from the teledata database
teledata-company:775 rdf:type owl:Company.
teledata-company:775 rdfs:label "Carl Zeiss Vision AG".
teledata-company:775 rdfs:label "American Optical Company International AG".
teledata-company:775 dbpedia-owl:numberOfEmployees "35".
teledata-company:775 dbprop-de:umsatz "4183400.0".
teledata-company:775 ofwi:company-status "active".
teledata-company:775 dbpedia-owl:industry ofwi-industry:8962, ofwi-industry:7752.
teledata-company:037041 schema-org:address ofwi-address:037041.

# link to the kompass database and kompass data
ofwi-company:037041 owl:sameAs teledata-company:775.
ofwi-company:037041 rdf:type dbpedia-owl:Company.
ofwi-company:037041 rdfs:label "Carl Zeiss Vision Swiss AG"@de.
ofwi-company:037041 dbpedia-owl:abstract "Zweck der Gesellschaft ist der Vertrieb..."@de.
ofwi-company:037041 schema-org:tickerSymbol "AFX.F".
ofwi-company:037041 dbprop:symbol "CZ" .

# contact information
ofwi-company:037041 dbprop-de:unternehmensform dbpedia-de:Aktiengesellschaft.
ofwi-company:037041 schema-org:email "office@zeis.com".
ofwi-company:037041 schema-org:faxNumber "055254473730".
ofwi-company:037041 schema-org:telephone "0552547373".
ofwi-company:037041 schema-org:email "info.swiss@vision.zeiss.com".
ofwi-company:037041 schema-org:url "http://www.vision.zeiss.ch".

# products and services
ofwi-company:037041 dbprop:products ofwi-productgroup:38371, ofwi-productgroup:3837122.
ofwi-company:037041 dbprop:distributor "Teflon easycare", "i.Profiler", "Carl Zeiss".
ofwi-productgroup:38371   rdfs:label "Optische Linsen", "Glaeser", "Spiegel".
ofwi-productgroup:3837122 rdfs:label "Brillenglaeser".

# key people
ofwi-company:037041 dbprop:keyPeople ofwi-person:Peter_Daepp_(037041);
                    dbprop:keyPeople ofwi-person:Sven_Hermann_(037041).

# address information
ofwi-address:037041 rdf:type schema-org:PostalAddress
ofwi-address:037041 schema-org:addressCountry "CH".
ofwi-address:037041 schema-org:addressRegion "ZH".
ofwi-address:037041 schema-org:postalCode "8714".
ofwi-address:037041 schema-org:addressLocality "Feldbach".
ofwi-address:037041 schema-org:streetAddress "Feldbacherstrasse 81".

# industry mapping
ofwi-industry:8962
   rdf:label "Wholesale of photographic and cinematographic equipment, precision ..."@en;
   rdf:label "Commercia all'ingrosso di apparecchi fotografici e cinematografic, ..."@it;
   rdf:label "Commerce de gros d'appareils photographicques et cin...."@fr;
   rdf:label "Grosshandel mit Foto- und Kinogeraeten, freinmechanische und optische ..."@de.
...
```

Figure 2: Excerpt of the company entry for the "Carl Zeiss Vision AG".

3.2 Linked open data

Recognyze disambiguates named entities by matching them to public and enterprise linked open data repositories. The system uses abstract SPARQL query profiles for mapping structured data retrieved from SPARQL endpoints to disambiguator classes. This generic approach allows using any structured data source that is accessible over SPARQL. Currently, query profiles for well-known sources include DBpedia[2] for identifying people and organizations, and GeoNames[6] for recognizing geographic locations. Future work will also explore options for combining linked enterprise data with open data sources to obtain further background information on locations, products and companies.

4 Method

The Recognyze component introduced in this paper identifies named entities in unstructured documents of heterogeneous origin (Recognyze currently accepts plain text and XML documents), and links these entities to structured knowledge repositories. We first describe obstacles towards using such data for text mining (Section 4.1) and then elaborate on how these repositories are leveraged in the disambiguation and named entity linking process which consists of the following two main tasks: (i) *linked data pre-processing* (Section 4.2) to extract search terms for identifying named entities, contextual information, and structural information from linked data sources, and (ii) *disambiguation and ranking* (Section 4.3) to locate and disambiguate these search terms in text documents.

4.1 Major challenges

This section introduces the following terminology for the purpose of discussing major challenges towards using linked enterprise data for named entity linking which will be used throughout the remainder of the article:

1. the *legal company name* refers to a company's official name, such as "International Business Machines Corporation" for IBM.

2. *search terms* or *search needles* are names used to identify mentions to a named entity in text documents, often derived from legal company names.

3. *ambiguous search terms* are search needles that are identical with commonly used words such as "Apple" or "international", which often do not refer to a named entity.

4. *unambiguous search terms* are considered specific enough to prevent ambiguities with common words (although they still may be used for multiple companies and therefore require a disambiguation step to distinguish these entities).

5. *candidate mentions* are mentions of an ambiguous or unambiguous search term in the document. These mentions may refer to a named entity in the knowledge base (Apple Inc.) or may proof to be unrelated to the data source (apple tree, apple juice, etc.).

[6] www.geonames.org

Table 3: Data pre-processing challenges

ID	description	Further information and examples
1	*data quality*	
1.1.	ambiguous short names	The knowledge source sometimes contains *short forms* of company names that are highly ambiguous. Example: Aktien Gesellschaft (English: incorporation), Hell (English: bright), Maximum (English: maximum), etc.
1.2.	uppercase only company names	use of uppercase only company names; these names are hard to find in regular documents and complicate the detection of abbreviations such as DER SA, DER HEIZER, DER ROTE SCHUH, etc.
2	*ambiguities*	
2.1.	many very small companies occur in the data set	A search for companies which include the name "Meyer" yields more than 1300 results in the raw data set. 1437 names contain the text "Personalfürsorgestiftung" and 1018 the term "Personalvorsorgestiftung".
2.2.	ambiguous company names	The problem of ambiguous company names is further complicated by the high level of granularity. For instance, Recognyze's knowledge base knows 13 different companies with the name "IST". The German Wikipedia, in contrast, does not contain a single company entry, referring to this name.
2.3.	legally related companies	Recognyze's knowledge base distinguishes 83 different legal entities with the name "Credit Suisse" and 92 entities which contain the name UBS. In contrast, Wikipedia contains only one entity for both companies.
2.4.	similar company names with no or little metadata	Some company entries consist of nearly identical names (e.g. ABSOLUT, ABSOLUT SA, ABSOLUT COSMETICS, etc.) and no or only little metadata which make it even for a human expert impossible to distinguish these name variants.
3	*low data granularity*	
3.1.	ambiguous company names	Company names such as IST (English: is), WEG (way)
3.2.	ambiguous person names	e.g. Robert Frey versus Robert Frey Consulting.
4	*use of casual name forms*	
4.1.	short names	Web pages often contain a company's short form rather than its legal name. Collaborative knowledge sources such as Wikipedia are more likely to include such forms. Example: "IST AG" rather than "Innovative Sensor Technology IST AG".
4.2.	use of "insider" casual names	Web pages use short name forms, that are not directly derived from the company's official name. Example: Sonova to refer to the Phonak Sounds AG, or CS is commonly used for Credit Suisse

The obtained linked enterprise data considerably differs from publicly maintained resources such as DBpedia, Freebase and Geonames:

1. it contains highly standardized data composed of *legal* company names and optional information on a company's address, management, and business areas. Depending on the source, different representations are used to express these data. Some sources only contain uppercase company names, for example, others tend to include shorter and often informal variations of the name.

2. the number of companies is considerably higher than in public sources, because the data set includes very small companies. The German version of Wikipedia, for instance, lists three companies with the ambiguous name "Total" as of September 2014. In contrast, the OFWI linked enterprise data repository contained 28 companies in business areas such as consulting, furniture, office management, fire protection equipment, vehicle halls, recycling and crude oil processing.

3. the enterprise data repository also contains historical company names which have proven to be another source for potential ambiguities.

The named entity linking requires search terms (search needles) to identify potential candidate named entities. A key issue when developing Recognyze, therefore, was enabling its data pre-processing components to automatically detect ambiguous company names and generate short name variations - unique to prevent ambiguities, yet short enough to be found in Web documents.

Table 3 summarizes the major obstacles towards generating unambiguous search needles for named entity linking from linked enterprise data. The following sections describe Recognyze's system architecture and provide a detailed description of how its components address the outlined challenges.

4.2 Linked data pre-processing

Figure 3 shows how Recognyze processes statements retrieved from linked data repositories to assemble disambiguation profiles that are then used by the named entity linking component. Recognyze uses application-specific profiles (e.g. `geonames_locations.en10000` for English location names of cities with a population of more than 10 000 inhabitants, or `ofwi_organizations_de` for German organization names). These profiles are stored in the *Recognyze configuration repository* and contain SPARQL queries that retrieve (i) the raw names (e.g., legal company names) of the entities, (ii) structural information, and (iii) context information such as products and services offered by an organization from the linked data repository as well as the mapping of these data onto the corresponding classes, pre-processors, and disambiguation algorithms. Analyzers then pre-process and assign this information to three different entity types - locations, people and organizations. Each entity contains at least a name field, a list of unique alternative names (e.g. abbreviations, colloquial names used to refer to the company and stock ticker symbols), a list of context terms such as products offered by companies, a company's address or management, a field indicating the entity's relative importance (e.g. based on its populations, revenue, or the number of citations), and data fields for encoding structural information.

Recognyze distinguishes three different analyzer types which populate these fields - name analyzers, structure analyzers and context analyzers. The raw company names available in the SPARQL repository roughly correspond to the names stored in the official company register. Although names such as "Credit Suisse Loan Funding LLC" are used in legal documents, they rarely occur in documents relevant for Web Intelligence such as news articles, product forums and social media sites. Recognyze therefore includes *name analyzers* which decompose legal company names obtained from the repository into ambiguous and unambiguous search needles that resemble the most probable names used to refer to the company.

The *structure analyzers* are used to extract and integrate structural and hierarchical information into the named entity linking process. The GeoNames repository, for instance, contains comprehensive information on geographic entities and their relations to each other. This allows deducing in which state and country a particular city is located, and provides information on nearby locations. Recognyze extracts comprehensive information on the relations between companies and their management from the enterprise linked data repository, which is then used to disambiguate companies which yield identical search needles.

Context analyzers handle context information obtained from the SPARQL queries. This information may yield additional context terms that have been generated from address information, products

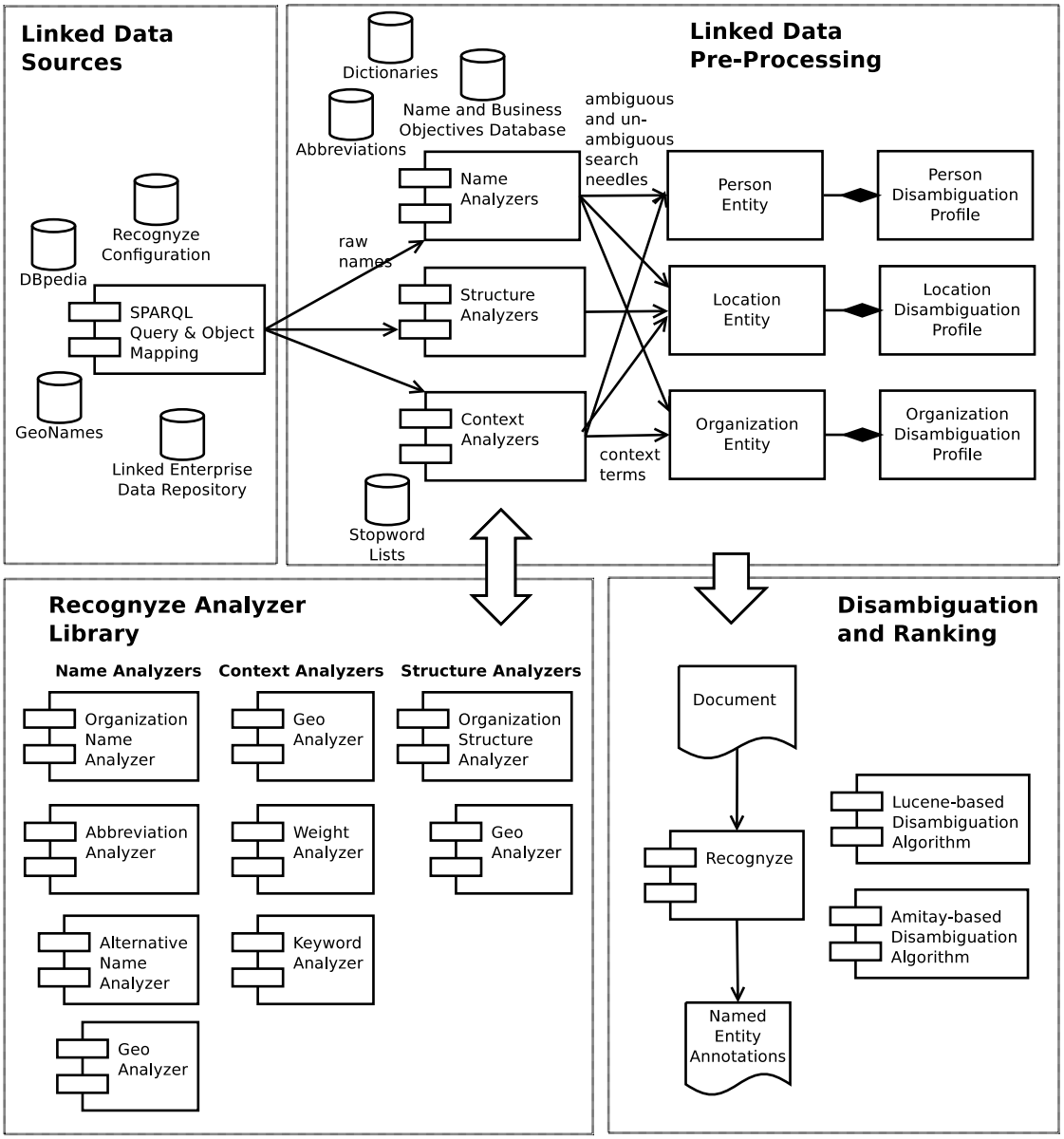

Figure 3: Named entity linking with Recognyze.

and services offered by a company, or numerical data such as a company's revenue and the number of employees that are then used as a weight in the disambiguation process (companies with higher revenue are considered more important than smaller companies).

4.2.1 Analyzers

Currently Recognyze supports the following pre-processing analyzers (Figure 3) which extract names, context information and structural information from data obtained from SPARQL queries.

- The **Abbreviation Analyzer** provides name variants and context information by extracting abbreviations from SPARQL query results. The profile can specify a minimum abbreviation length, whether normalized variants (e.g., OMV for ÖMV) should be generated, and stopword lists to prevent the inclusion of common abbreviations such as currency codes. This analyzer has proven especially useful for automatically extracting stock ticker symbols and abbreviations from `dbpedia-owl:abstract` properties.
- The **Alternative Name Analyzer** extracts name variants for person or organization objects and performs only a minimum pre-processing such as automatically converting URLs

which include an organization name to the corresponding name. The URL `dbpedia.org/page-/Facebook,_Inc`, for instance, is translated into the string `Facebook, Inc`.

- **Geo Analyzer** extracts name, context and structural information from the data fields name, alternative name, population, parent, latitude, longitude, altitude, feature code, parent country and official name, and maps these fields to the corresponding attributes of geo objects. The name and official name bindings yield the entity's official name in the chosen language, the feature code provides context information on the entity type (e.g. city, district, state, country, etc.), parent and parent country reveal where the entity is located (e.g. Zurich in Switzerland) and are used for constructing hierarchical entity trees, which are then used for disambiguation. The population field provides information for ranking search results, i.e. entities which have a larger population are preferred over smaller entities.

- The **Keywords Analyzer** extracts context information for person and organization objects. The analyzer allows specifying a minimum token length, whether normalized keywords variants (e.g. Baloise for Bâloise) should be generated, and whether lowercase keywords should be ignored.

- **Organization Name Analyzer** is the most advanced analyzer for extracting organization names from SPARQL bindings. It uses an entropy-based heuristic to create search needles from the raw company name obtained from the knowledge source that are (i) short enough to occur in informal textual documents such as News articles, and (ii) unique enough to prevent ambiguities. The following pseudo code outlines how the Organization Name Analyzer evaluates the ambiguity of an organization name. The algorithm first splits the company name into tokens (line 7) and then assesses the entropy contributed by each token.

```
1:  isCaseSensitive ← isCaseSensitive(companyName)
2:  if isCaseSensitive then
3:      nameEntropy ← 0
4:  else
5:      nameEntropy ← -0.25
6:  end if
7:  companyNameTokens ← split(companyName)
8:  lastToken ← ∅
9:  for all token in companyNameTokens do
10:     nameEntropy += getEntropy(token, tokenPos, isCaseSensitive)
11:     lastToken = token
12: end for
13: classEntropy ← getClassEntropy(companyNameTokens)
14: if (isConnector(lastToken) || isPossessive(lastToken)) then
15:     return INVALID_NAME
16: else if (isCommonTerm(companyName)
        || isGeoName(companyName)
        || isPersonName(companyName)
        || len(companyName) < MIN_NAME_LEN
        || nameEntropy + classEntropy < 1.1 − len(companyNameTokens) ∗ 0.1) then
17:     return AMBIGUOUS_NAME
18: else
19:     return UNAMBIGUOUS_NAME
20: end if
```

The component's *getEntropy* function (line 10) has access to (i) lexicons of common Swiss, Italian, German and French last names, given names, business objectives, company types and abbreviations (Section 4.2.2), and to (ii) heuristic rules which determine whether a word is likely to be part of an abbreviation (e.g. IBM versus BIOTEC), a possessive form (e.g. Swiss), or a connector - which would require an additional token in the company name to assess the entropy contributed by a single token. In addition, it considers information on the number of different name component classes (e.g. abbreviation, name, dictionary term, trade) used in the company name (line 13). Our experiments have shown that names consisting of components from different classes have a higher probability of being unique than names with a lower number of classes (e.g.

only names). The name analyzer, therefore, awards extra entropy for every name component class included in the final company name.

Another issue to address are case insensitive names (compare Table 3, Challenge 1.2.). The name analyzer penalizes case insensitive names with negative initial entropy and by switching to case insensitive look-ups for the token classification (line 2). Therefore, case insensitive names need to include more tokens, before name analyzer considers them unambiguous search terms. The needles returned by the name analyzer satisfy the following three criteria:

1. they contain of at least three characters and exceed a minimum entropy threshold. The entropy threshold ensures, that the names are unique enough to prevent ambiguities with common terminology and phrases,
2. they do not end with a connector or possessive form. The names are complete enough to be recognized as full company names - this prevents broken names such as "Zingg &" or "Gesellschaft Schweizerischer" (Society of Swiss), which are reported as invalid names (line 15).
3. they are not identical to common terms found in an English, French or German dictionary and do not consist of a single first or last name.

Table 4 illustrates example mappings that have been derived using this method. Needles that do not satisfy these criteria are considered ambiguous and, therefore, require a special treatment in the disambiguation component.

Table 4: Examples for automatic mappings of legal company names to search needles.

Legal company name	**Search needle**
Atelier Architrav Baumann Rolf Architekt HTL	Atelier Architrav Baumann
Crédit Suisse AG	Crédit Suisse
IBM (Schweiz), Zweigniederlassung Basel	IBM
IBM Research GmbH	IBM Research
OK Coop Tankstelle Vaduz GmbH, mit Sitz in Kriens	OK Coop Tankstelle
Restaurant Coop L'Aidjolat, Bruno Migy	Restaurant Coop
Zingg & Nüssli, Architekt und Ingenieur	Zingg & Nüssli

- **Organization Structure Analyzer** extracts structural information such as relations between organizations (subsidiary, competitor, etc.) and relations between organizations and their management.
- **Simple Name Analyzer** is a simple domain-independent component which extracts company names from SPARQL bindings and stores them in the corresponding name and alternative name fields without any further pre-processing.
- The **Weight Analyzer** inspects fields such as the number of employees, a company's revenue or an geographic entity's population to obtain information on the entity's importance which is then stored in a `weight` attribute.

4.2.2 Lexicons

Lexicons complement analyzers by providing them with language specific background information on common abbreviations, location names, person names, products, and stopwords relevant for a particular language. Based on these lexicons analyzers are able to tag ambiguous names, to estimate the likelihood of name variants, to identify frequently used person names, and to disambiguate terms due to their use together with defined prefixes or suffixes (Section 4.3). Table 5 lists lexicons which are frequently used in the pre-processing configurations.

Table 5: Lexicons frequently used in Recognyze pre-processing configurations.

Lexicon Group	Lexicon	Description
abbreviations	business	commonly used abbreviations in business documents
	currencies	international currency codes
	political parties	country and language dependent abbreviations of major political parties
dictionaries	-	dictionaries in different languages.
locations	popular	popular location names such as Wall Street and Downing Street
	countries	a list of all country names
organizations	Forbes	all companies listed in the Forbes Global 2000
	prefix	typical prefixes which indicate organizations
	suffix	typical suffixes indicating organizations
people	firstnames	country and language specific lists of the 150 most popular first names
	powerful	names of the most powerful people according to Forbes
products	popular	the most popular product names
stopwords	-	language-specific stopwords

4.2.3 Recognyze profile configuration

Recognyze uses profile configurations which (i) specify SPARQL queries which acquire relevant data field, and (ii) how these fields are mapped and pre-processed to extract entities which are then used in the disambiguation process.

Every Recognyze profile configuration consists of the following components:

1. profile metadata describing the profile name and the SPARQL server on which the query is processed (queries may also spawn multiple SPARQL servers by using `SERVICE` statements in the query).

2. a SPARQL query which defines bindings (i.e. data fields) for evaluation by Recognyze's data pre-processing pipeline.

3. analyzers for processing the data obtained by the query. Analyzers process at least one binding and bindings may be used by multiple analyzers. The `company_name` field, for instance, may be used by the Organization Name Analyzer to generate name variants *and* by the Keyword Analyzer to obtain keywords which provide additional context information.

The specification of lexicons allows Recognyze to detect ambiguous names, to apply disambiguation techniques (Section 4.3) and to use advanced analyzers such as Organization Name Organizer which perform more complex name analyses. Table 6 presents an excerpt of a Recognyze profile.

Figure 4 illustrates how name analyzers pre-process SPARQL query results based on the profile configuration outlined in Table 6 and the example company entry for the *Carl Zeiss Vision AG* (Figure 2). The Organization Name Analyzer generates unambiguous name variants from the official company names found in the `company_name` column of the SPARQL result. The lexicons let the analyzer assess name variants and decide on whether additional tokens are required for generating an unambiguous company name.

The Keyword Analyzer extracts context information from the `abstract`, `company_city`, `key_person_firstname`, `key_person_lastname`, `company_products` and `company_industry` columns, the Alternative Name Analyzer provides further name variants based on the `company_url` and `ticker_name` fields, and the Weight Analyzer computes the company's importance based on its `turnover_weight`. The obtained *Organization Entity* is then used for disambiguation and ranking if one of its names is mentioned in a document.

Table 6: Simplified excerpt of a Recognyze profile.

field	value
label	`eval.5.context`
source	`http://.../r/lod.ofwi.ch`

query	SELECT ?s ?company_name ?abstract ?key_person_firstname ?key_person_lastname ?company_url ?company_city ?company_country ?company_products ?company_industry ?turnover_weight ?ticker_name
	WHERE {
	?s rdf:type dbpedia-owl:Company.
	?s rdfs:label ?company_name .
	OPTIONAL { ?s schema-org:address ?ofwi_address .}
	...
	FILTER (LANG(?company_name) = "de")
	...
	}

entity type	`Recognyze.OrganizationEntity`
lexicon	`scope: name, dict=dict.C, dict.de, dict.en,`
lexicon	`scope: firstnames, dict=firstname.de, firstname.en, ...`
affix lexicon	`Recognyze.OrganizationAffix`
analyzer	`nameContext`
analyzer	`abbreviation`

4.3 Disambiguation and ranking

The *Recognyze disambiguation process* draws upon the disambiguation profiles created by pre-processing of the knowledge base (Figure 3). Agents that call Recognyze have to specify the incoming documents and disambiguation profiles to be applied in the named entity linking process. To identify candidate mentions (names and alternative names) and the corresponding context information, the component then searches every document for occurrences of

- the unambiguous search needles that have been generated by the name analyzers,
- the ambiguous search terms which are either *prefixed* or *suffixed* by terms that indicates that they refer to a company. Typical prefixes are trades (Firma/company, Hotel/hotel, Gasthaus/restaurant) while terms indicating a company's legal status such as AG/Inc, GmbH/Limited, are used as suffixes, and
- structural information and context terms which are then used to disambiguate companies with identical search terms. This step is particular important since the linked enterprise data repository comprises a significantly higher number of companies than publicly available data sources.

To identify specific entities, the system then uses a profile-specific disambiguation algorithm such as Amitay[1] for locations, or an adapted version of the Lucene similarity search described in Equation 1 for people and organizations.

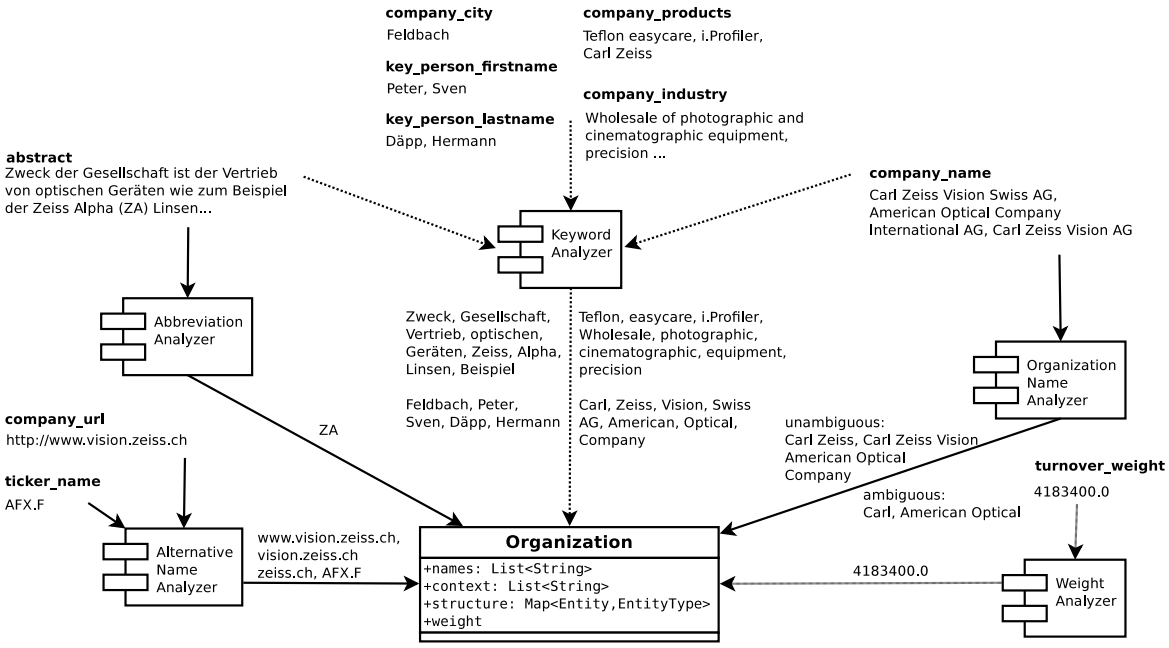

Figure 4: Data extracted from SPARQL results by the Name Analyzer of Recognyze - solid lines indicate names, dashed lines context information and dotted lines weights.

$$s(q_e, d) \quad = \quad coord(q_e, d) \cdot |q_e| \sum_{t \in q_e} \left[idf(t)^2 \cdot boost(t) \right] \tag{1}$$

The algorithm determines an entity's (e) significance ($s(q_e, d)$) in document (d) based on the occurrences (q_e) of the entity's unambiguous and ambiguous search terms, and its context terms obtained from pre-processing. The inverse document frequency ($idf(t)$) value ensures that rare search terms provide a higher contribution to the total score. The coordination factor ($coord(q_e, d)$) rewards entities which match multiple times, for example when relevant context terms are mentioned in addition to the company name. The boost factor ($boost(t)$) is used for modifying the significance score to influence the selection of entities - to favor longer matches over shorter ones, for example.

Recognyze computes the boost factors $boost(t)$ based on the needles' source and the number of times it appears in the document. Full matches of a (short) company name or an alternative company name obtain high boost factors, while matches of context terms yield considerably lower boost factors. The entities are then ranked according to their score ($s(q_e, d)$). In cases where multiple entities obtain the same score, Recognyze uses further structural and context information such as the company's revenue and its number of employees to finalize the ranking.

Recognyze's default setting tends to return duplicate entities - i.e., different branches and subsidiaries of a company which has been identified with high confidence (these duplicate entities obtain high confidence values due to the needles contributed by the high-confidence company). To prevent such duplicates and return more heterogeneous and useful results, entities can be re-scored to preserve other entities in the document. When iterating through the set of results, the re-scoring algorithm keeps the most significant entity and removes the corresponding needles from the evaluation. This removes the bias which leads to the inclusion of duplicates.

For the named entity linking of geographical entities, structural relations between geographic locations can support the disambiguation process - e.g., a reference to Vienna is more likely to refer to Vienna/Austria than to Vienna/Massachusetts if the entity Salzburg/Austria is mentioned in the same document. Future versions of Recognyze will apply this disambiguation technique to the identification of people and organizations as well.

5 Evaluation

The algorithms used for identifying locations have been thoroughly described in earlier work[29]. Therefore, this section will focus on organizations and assess whether Recognyze provides an accurate and scalable named linking component for this entity type. Future work will extend the evaluation to additional entity types such as people and products.

Since the linked data repository often is restricted to a company's legal name (e.g. Crédit Suisse AG), but does not contain frequently used abbreviations such as CS and stock ticker symbols, we manually extended the linked enterprise knowledge source with these entries for all companies listed in the Swiss Market Index (SMI). If DBpedia is used as data source, Recognyze's Abbreviation Analyzer automatically extracts such abbreviations from the entity's description.

The detection of organizations is a challenging task due to the enormous amount of background information yielded by a repository of more than 2.9 million companies that need to be considered in the disambiguation step. Iterative optimizations helped to improve throughput and memory consumption of Recognyze.

5.1 Data sources

The evaluation has been performed on the following data sets:

1. The *AWP.ch financial news* data set provided by OFWI is stored in a 260 MB CSV file with more than 320,000 news messages. Each message contains a company id that corresponds to the identifiers used in the linked enterprise data repository, the company name, a unique message id, timestamp, message source, topic, language, title and message content.

   The evaluation component uses the company id for verifying whether Recognyze has been able to correctly identify the company based on the message content. The experiment uses a randomly selected subset of German-speaking news messages that were annotated with *exactly one* company which is supposed to be the predominant named entity in that particular document. The resulting test corpus contains a total of 50 000 document with 1 175 different companies and organizations. The goal of this evaluation is to (i) determine how well Recognyze is able to identify organizations within this data set, and (ii) how well the ranking of Recognyze's scoring algorithm corresponds to the ranking of human experts regarding the most relevant company for a particular document.

2. An *extended AWP business news data set* which consists of 150 randomly selected German-speaking news messages that have been manually annotated by domain experts. The annotations cover *all* companies in a particular document.

3. The *NZZ (Neue Zürcher Zeitung; New Journal of Zurich) news data set* was compiled out of 150 randomly selected NZZ business news articles, which were published between 1 August and 30 September 2013 (human evaluators annotated all named entities in these articles).

We use the last two evaluation data sets to contrast Recognyze's named entity linking performance for documents from rather formal business news (AWP data set), as compared to documents from less formal newspaper articles (NZZ data set). The latter cover a much larger range of topics and are, therefore, expected to be more prone to ambiguities.

5.2 Evaluation settings

The evaluation has been designed to demonstrate the impact of the following three factors on the named entity linking and ranking performance:

1. the pre-processing of raw names which deals with the trade-off between preventing ambiguities (high precision) and high coverage of all variants of company names (high recall). The evaluation contrasts the following five name pre-processing strategies: (i) *raw names* uses the names of the knowledge source without any pre-processing; (ii) *simple* tokenizes names and transfers them into a standardized form, (iii) *simple & filtering* performs simple pre-processing and then removes needles which are composed of stopwords or dictionary items; (iv) *advanced* uses Organization Name Analyzer (Section 4.2.1) for the name pre-processing, and (v) *advanced*

& filtering performs the advanced name pre-processing and a filtering step for needles composed of dictionary terms.

Table 7 outlines the impact of the chosen name processing strategy on the terms that Recognyze considers valid company names (which will be identified as companies) for the example data set from Figure 2. The *raw names* pre-processing strategy, which is used by most state-of-the-art named entity linking techniques, only considers verbatim matches. *Simple* pre-processing, by contrast, considers all mono- and n-grams that were extracted from the knowledge base as potential names. *Advanced* pre-processing balances both strategies by evaluating the entropy of potential name candidates (Section 4.2.1) to identify unambiguous name candidates, and requiring prefix or suffix indicators in the case of ambiguous name candidates (Section 4.3). Filtering removes name candidates that are also found in dictionaries (Section 5) and are, therefore, considered ambiguous. In the example presented in Table 7, filtering does not affect the results for the advanced name processing, since no dictionary terms have been suggested by the Organization Name Analyzer component.

Table 7: Impact of the chosen processing strategy on the selection of company name candidates.

Processing Strategy	**Company Name Candidates**
Raw names	www.vision.zeiss.ch, AFX.F, Carl Zeiss Vision Swiss AG, American Optical Company International AG, Carl Zeiss Vision AG
Simple	www.vision.zeiss.ch, AFX.F, Carl, Zeiss, Vision, Swiss, AG, American, Optical, Company, International, Carl Zeiss, Zeiss Vision, American Optical, Optical Company
Simple & filtering	www.vision.zeiss.ch, AFX.F, Carl, Zeiss, Carl Zeiss, Zeiss Vision, American Optical, Optial Company, ...
Advanced	*unambiguous:* Carl Zeiss, Carl Zeiss Vision, American Optical Company, www.vision.zeiss.ch, vision.zeiss.ch, zeiss.ch, AFX.F; *ambiguous:* Carl, American Optical
Advanced & filtering	*unambiguous:* Carl Zeiss, Carl Zeiss Vision, American Optical Company, www.vision.zeiss.ch, vision.zeiss.ch, zeiss.ch, AFX.F; *ambiguous:* Carl, American Optical

2. to which extend Recognyze considers context information (i.e. whether we disambiguate ambiguities between organizations based on the context terms yielded by the name analyzers), and

3. the strategy used for ranking articles, with or without re-scoring. Please refer to Section 4.3 for a more detailed description of the re-scoring algorithm.

Recognyze uses context information for disambiguation and entity ranking if the context flag is set. Context information for the disambiguation of organizations comprises information on the company's management, address, products and the industry sector the company operates in. The named entity ranking algorithm also considers information on the company's revenues and the number of employees.

5.3 Normalization

The enterprise data repository contains a fine-grained description of legal entities. For instance, there are ten different branch offices of the company HG Commerciale, a Swiss provider for building materials, listed in the database, and more than 100 different branches and subsidiaries of the UBS

bank. Distinguishing such entries from each other is outside the scope of Recognyze and of most human experts. Therefore, we normalize closely related entities by merging them into a single entity prior to comparing Recognyze's output to the gold standard.

A data pre-processing module maps such legally related entities onto the company with the highest reported revenue, and retrieves data on company agglomerates and ownership structures to identify cases where an article has been assigned to a parent company rather than to the company mentioned in the article.

The following pseudo code illustrates how the algorithm pools companies that share the same *commonPrefix* to a single entity.

```
1:  commonPrefix ← ''
2:  tokenPos ← 0
3:  for all word in companyName do
4:     commonPrefix ← commonPrefix + word + '␣'
5:     if NOT isIgnoreTerm(word) then
6:        tokenPos ← tokenPos + 1
7:     end if
8:     if NOT (isAbbreviation(word) OR isName(word)
           OR isCommonTerm(word, tokenPos)) then
9:        return commonPrefix
10:    end if
11: end for
12: return commonPrefix
```

This prefix is computed by assembling words that are sufficient to distinguish the company from other (unrelated) organizations. The algorithm, therefore, requires additional tokens for words that either contain typical French, German, or Italian names (`isName`), terms commonly used in Swiss company names such as AG, Suisse, GmbH (`isCommonTerm`), one letter abbreviations (`isAbbreviation`) or irrelevant terms such as prepositions (`isIgnoredTerm`). The evaluation also uses the word position for evaluating, whether a word is considered a common term or not.

The company mapping performed by the algorithm has been verified by two independent domain experts prior to the evaluation step.

5.4 Results

Table 8 summarizes the performance of Recognyze's named entity ranking - i.e., how well the most significant named entity returned by Recognyze correspond to the preferences of the domain experts who assigned exactly one company to each of the 50,000 evaluated articles. Recognyze's recall of the domain experts evaluation (R@1) indicates that raw names yield a maximum recall of 0.69. Name pre-processing performs best in this setting, since it generates name variants which correspond well to the names used in formal business news. Considering the message context further improves the component's performance.

Table 8: Recognyze named entity linking and ranking performance on the full AWP data set.

name processing	**context**	**R@1**
Raw names	·	0.60
	✓	0.69
Simple	·	0.56
	✓	0.53
Simple & filtering	·	0.62
	✓	0.72
Advanced	·	0.65
	✓	0.73
Advanced & filtering	·	0.68
	✓	0.72

Table 9 illustrates the influence of the domain on the use of company names. The evaluation is based on the manually annotated set of 150 NZZ Newspaper articles and 150 AWP messages, and uses Recognyze setting which maximizes recall. The recall value provides an indication for the coverage of the named entity knowledge base and establishes an upper boundary of Recognyze's recall with the current name pre-processing. For instance, since Newspaper articles tend to use informal company names (such as IBM rather than IBM Switzerland AG), the coverage provided by raw names obtained from the linked enterprise database is comparably low. The AWP business news messages are not that much affected, since the use of formal company names is much more common in this setting.

Table 9: Estimated coverage of the named entity knowledge base.

name processing	**rescore**	AWP messages **R**	NZZ articles **R**
Raw names	·	0.52	0.13
	✓	0.52	0.13
Simple	·	0.95	0.95
	✓	0.81	0.66
Simple & filtering	·	0.87	0.71
	✓	0.78	0.55
Advanced	·	0.88	0.82
	✓	0.84	0.78
Advanced & filtering	·	0.87	0.81
	✓	0.83	0.76

Applying the pre-processing techniques discussed in Section 4 significantly improves the coverage of entity names. This is especially true for the simple pre-processing which generates tokens composed of the original company names and, therefore, provides the highest recall. Such a high recall comes at a price - many false positives and a much lower performance if the balance between precision and recall is taken into consideration as demonstrated in the next evaluation.

Table 10 summarizes the results of the named entity linking. Again, there is a clear correlation between the applied name pre-processing and the obtained performance. Evaluations which use the raw names (no name pre-processing) or only a simple pre-processing obtain significantly lower results than evaluations that apply the advanced pre-processing techniques. This is especially true in less formal settings such as Newspaper articles, where raw names obtain a recall as low as 0.13. Simple pre-processing considerably improves this number for Newspaper articles but at the cost of a very low precision due to ambiguous needles. The filtering of ambiguous terms improves overall performance, although it still remains too low to obtain usable results.

Applying the advanced name pre-processing capabilities offered by the Organization Name Analyzer considerably improves precision and recall in all settings. If name analyzer is combined with filtering we obtain a recall of 0.80 (0.74) for AWP (NZZ) articles and an F1 measure of 0.59 (0.63). Table 10 also shows that contextualization needs a minimum quality of the search needles to be effective. For that reason, contextualization only yields significant improvements for the advanced name pre-processing.

This observation is also true for re-scoring, which is not effective for raw names and the simple pre-processing, but significantly improves results ones the needle quality is appropriate.

5.5 Discussion

The results presented in the previous section demonstrate how the progression from simple to more advanced name pre-processing, disambiguation and filtering strategies improves the performance of named entity linking. A qualitative analysis of incorrectly classified documents identified the following most prominent reasons for failed named entity linking attempts:

1. *ambiguous company names:* the Organization Name Analyzer marked the company name as ambiguous and the text only contained the ambiguous name without any of the prefixes or

Table 10: Recognyze named entity linking performance on the extended NZZ and AWP data sets.

name			AWP messages			NZZ articles		
processing	**context**	**rescore**	**P**	**R**	**F1**	**P**	**R**	**F1**
Raw names	·	·	0.44	0.52	0.44	0.14	0.13	0.11
	·	✓	0.48	0.52	0.46	0.16	0.13	0.12
	✓	·	0.46	0.52	0.44	0.14	0.13	0.11
	✓	✓	0.49	0.52	0.47	0.16	0.13	0.13
Simple	·	·	0.07	0.52	0.10	0.03	0.45	0.06
	·	✓	0.07	0.65	0.12	0.04	0.58	0.07
	✓	·	0.06	0.48	0.09	0.03	0.36	0.05
	✓	✓	0.09	0.61	0.14	0.04	0.55	0.07
Simple &	·	·	0.15	0.62	0.19	0.07	0.50	0.11
filtering	·	✓	0.24	0.76	0.34	0.15	0.55	0.22
	✓	·	0.15	0.67	0.21	0.07	0.54	0.11
	✓	✓	0.26	0.78	0.36	0.16	0.58	0.24
Advanced	·	·	0.32	0.71	0.38	0.28	0.74	0.37
	·	✓	0.34	0.84	0.45	0.35	0.78	0.44
	✓	·	0.34	0.78	0.43	0.29	0.76	0.38
	✓	✓	0.35	0.83	0.46	0.37	0.78	0.46
Advanced	·	·	0.36	0.71	0.41	0.38	0.75	0.46
& filtering	·	✓	0.37	0.82	0.48	0.44	0.76	0.52
	✓	·	0.45	0.77	0.53	0.49	0.73	0.54
	✓	✓	0.50	0.80	0.59	0.60	0.74	0.63

suffixes required for disambiguation. An example would be mentions of "Die Post" (the post) which in German either refers to the company or to mail received.

2. *different spelling variants:* the document used a different spelling variant of the company name such as for example "Job Up" rather than the name "JobUp" which was recorded in the database.

3. *missing name variants or abbreviations:* the text used name variants or abbreviations which have not been included in the linked enterprise data repository. For instance, a company's official name is "Hottinger Züri Valore AG", name analyzer created the unambiguous short company name "Hottinger Züri" but "Hottinger Zürich" was used in the document. Another common problem which falls into this category are entities such as the "Waadtländer Kantonalbank (BCV)" where the German name is included in the repository but the French name (Banque Cantonale Vaudoise) is used in the text. A possible solution to this problem could be obtaining needles from all three language variants (German, French and Italian) present in the knowledge repository.

For the entity ranking task (compare Table 8), two additional error sources have been identified:

1. the company used to annotate the article has not been named in the text. Such cases may appear if the article focuses on a subsidiary rather than on the parent company and the relationship between the two companies has not yet been documented in the linked enterprise data repository.

2. the company has been mentioned in the document, but other companies that also occur in the text have been returned by the tagger. We have limited the evaluation to documents annotated with only one company. Nevertheless, an analysis of documents that had been "incorrectly" classified revealed that some of these documents contain multiple organizations because they cover court cases, joint ventures, mergers and acquisitions. These examples demonstrate that even manually annotated and commonly used reference data sets contain a certain margin of error.

Comparing the obtained results to the literature is problematic since the reported accuracies strongly depend on the chosen test set and genre. Hachey et al.[10] present a comprehensive comparison of three different named entity linking approaches and return an accuracy between 77.6 and 80.8%

for the recognition of organizations in news entries and between 83.6 and 90.0% for Web pages on the NIST Text Analysis Conference (TAC) 2010 data set. Fernández et al.[7] report a *disambiguation* accuracy of 96% for their named entity disambiguation approach. This accuracy has been measured for the disambiguation process (but not for the overall named entity linking), requires a supervised learning algorithm and, therefore, feedback from human experts for adaptation to a particular domain. Evaluations that focus on an algorithm's disambiguation capacity (i.e. its capability to distinguish two ambiguous entities) rather than its total accuracy in regard to a labeled test corpus yield higher total accuracies because they do not need to consider cases where no valid entities have been found.

Generic methods do not achieve the accuracy of approaches which have been tailored to a specific domain, but provide the benefit of a relatively stable performance across different domains and settings. For this reason the evaluation used two rather extreme settings: (i) news articles using a rather informal language to refer to company names, and (ii) messages from the AWP business news service which tends to use the official company names. Since the evaluation is based on Swiss company names and news articles, French and Italian company names are frequently used in addition to German and English references.

It is important to note that the linked enterprise data repository used for evaluation purposes was much more fine grained than Wikipedia. For instance, it contained more than 83 different legal entities with the name "Credit Suisse" (versus one in the German Wikipedia as of September 2014) or 28 companies with the name "Absolut" (versus three in Wikipedia). Due to the vast amount of businesses registered in the database, it also contains highly ambiguous company names such as "sich bewusst sein" (to be aware of), "Die letzte Ruhe" (the final resting place), or "Der rote Schuh" (the red shoe).

In light of these challenges, Recognyze produced respectable results, especially when considering that it had not been adapted to the evaluation corpus (such a customization would defy generic applicability as one of the major design goals).

6 Outlook and Conclusions

This article presents Recognyze, a named entity linking component that uses background knowledge from linked enterprise data or linked open data such as DBpedia and GeoNames. In contrast to other approaches, Recognyze does not rely on machine learning and therefore does not require training corpora or iterative learning steps. Instead, it employs sophisticated data pre-processing modules to extract (i) ambiguous and unambiguous company names, (ii) context information and (iii) structural indicators from these sources. The extracted information is then used for linking and ranking named entities.

The transformation of heterogeneous enterprise data into linked enterprise repositories poses a number of challenges that are being addressed by Recognyze. The discussion of its analyzer modules illustrates the flexibility of the approach and demonstrates the potential of advanced pre-processing techniques to overcome shortcomings in the original data sets. The high recall for named entities referenced in business documents can be attributed to the use of a comprehensive linked enterprise repository, which contains detailed background knowledge to support the named entity linking process. The recall is lower when processing more informal sources such as news articles, but can be improved through the pre-processing methods introduced in this paper.

Although the literature has reported higher accuracies for certain entity linking approaches based on machine learning techniques, Recognyze is better suited for many real-world applications since it

1. is not limited to a particular knowledge source,
2. does not require training or annotated training corpora, but can be deployed for any domain or language as long as appropriate linked data resources such as DBpedia are available, and
3. offers good overall performance even with comprehensive knowledge bases such as linked enterprise repositories containing the full set of companies present in an official company directory, rather than the much smaller set of companies present in public knowledge sources such as DBpedia.
4. provides name analyzers to extract name candidates, context information, weightings and structural information from knowledge sources. These pre-processing capabilities are essential for

capturing entity mentions that do not exactly match the name string given in the knowledge source.

The evaluation supports the claim that Recognyze successfully disambiguates and grounds named entities in settings where a lot of similarly named alternatives (e.g. ambiguous company names such as "Total" or "Absolut") and collisions with common terms such as "to be aware of" occur. Depending on the used evaluation corpus, Recognyze yields a recall of 0.72 for identifying the most relevant organization in an article and an F1 measure of up to 0.63 for named entity linking, without data source-specific optimizations or human interventions.

Future work will focus on further improving Recognyze's disambiguation performance by considering more complex structural knowledge in the linking process. We will also optimize and evaluate disambiguation profiles that work with publicly available sources such as DBpedia, explore options for combining linked enterprise data with linked open data, provide evaluations for other entity types such as events and people, and extend our approach to French, Spanish and other languages.

Acknowledgment

The research presented in this paper has been conducted as part of the COMET Project (www.htw-chur.ch/comet), funded by the Swiss Commission for Technology and Innovation (CTI), and the DecarboNet project (decarbonet.eu), funded by the European Union's 7th Framework Programme for research, technology development and demonstration under the Grant Agreement No. 610829.

References

[1] E. Amitay, N. Har'El, R. Sivan, and A. Soffer. Web-a-where: geotagging web content. In *SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 273–280, New York, NY, USA, 2004. ACM.

[2] C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, and S. Hellmann. DBpedia - a crystallization point for the web of data. *Journal of Web Semantics: Science, Services and Agents on the World Wide Web*, 7(3):154–165, 2009.

[3] S. Chaudhuri, U. Dayal, and V. Narasayya. An overview of business intelligence technology. *Communications of the ACM*, 54(8):88–98, Aug. 2011.

[4] H. Chen. Business and market intelligence 2.0. *IEEE Intelligent Systems*, 25(1):68–83, 2010.

[5] J. Daiber, M. Jakob, C. Hokamp, and P. N. Mendes. Improving efficiency and accuracy in multilingual entity extraction. In *Proceedings of the 9th International Conference on Semantic Systems (I-SEMANTICS'13)*, page 121–124, 2013.

[6] A. Das and B. Gambäck. Sentimantics: conceptual spaces for lexical sentiment polarity representation with contextuality. In *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis*, WASSA '12, page 38–46, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

[7] N. Fernández, J. Arias Fisteus, L. Sánchez, and G. López. IdentityRank: named entity disambiguation in the news domain. *Expert Systems with Applications*, 39(10):9207–9221, 2012.

[8] J. R. Finkel, T. Grenager, and C. Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 363–370, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

[9] A. Gangemi. A comparison of knowledge extraction tools for the semantic web. In P. Cimiano, O. Corcho, V. Presutti, L. Hollink, and S. Rudolph, editors, *The Semantic Web: Semantics and Big Data*, number 7882 in Lecture Notes in Computer Science, pages 351–366. Springer Berlin Heidelberg, Jan. 2013.

[10] B. Hachey, W. Radford, J. Nothman, M. Honnibal, and J. R. Curran. Evaluating entity linking with wikipedia. *Artificial Intelligence*, 194:130–150, 2013.

[11] X. Han and J. Zhao. Named entity disambiguation by leveraging wikipedia semantic knowledge. In *Proceedings of the 18th ACM conference on Information and knowledge management*, CIKM '09, page 215–224, New York, NY, USA, 2009. ACM.

[12] S. Harris, T. Ilube, and M. Tuffield. Enterprise linked data as core business infrastructure. In D. Wood, editor, *Linking Enterprise Data*, pages 209–219. Springer US, Jan. 2010.

[13] J. Hoffart, Y. Altun, and G. Weikum. Discovering emerging entities with ambiguous names. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14, pages 385–396, Republic and Canton of Geneva, Switzerland, 2014. International World Wide Web Conferences Steering Committee.

[14] J. Hoffart, F. M. Suchanek, K. Berberich, and G. Weikum. YAGO2: A spatially and temporally enhanced knowledge base from wikipedia. *Artificial Intelligence*, 194:28–61, 2013.

[15] J. Hoffart, M. A. Yosef, I. Bordino, H. Fürstenau, M. Pinkal, M. Spaniol, B. Taneva, S. Thater, and G. Weikum. Robust disambiguation of named entities in text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, page 782–792, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[16] C. Hondros. Standardizing legal content with OWL and RDF. In D. Wood, editor, *Linking Enterprise Data*, pages 221–240. Springer US, Jan. 2010.

[17] J. J. Jung. Online named entity recognition method for microtexts in social networking services: A case study of twitter. *Expert Systems with Applications*, 39(9):8066–8070, 2012.

[18] S. S. Kataria, K. S. Kumar, R. R. Rastogi, P. Sen, and S. H. Sengamedu. Entity disambiguation with hierarchical topic models. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '11, page 1037–1045, New York, NY, USA, 2011. ACM.

[19] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. v. Kleef, S. Auer, and C. Bizer. DBpedia - a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web Journal*, 2014.

[20] S. Negash and P. Gray. Business intelligence. In F. Burstein, C. Holsapple, S. Negash, and P. Gray, editors, *Handbook on Decision Support Systems 2*, International Handbooks Information System, pages 175–193. Springer Berlin Heidelberg, 2008.

[21] J. Nothman, N. Ringland, W. Radford, T. Murphy, and J. R. Curran. Learning multilingual named entity recognition from wikipedia. *Artificial Intelligence*, 194:151–175, 2013.

[22] T. Omitola, J. Davies, A. Duke, H. Glaser, and N. Shadbolt. Linking social, open, and enterprise data. In *Proceedings of the 4th International Conference on Web Intelligence, Mining and Semantics (WIMS14)*, WIMS '14, pages 41:1–41:8, New York, NY, USA, 2014. ACM.

[23] A. Pilz and G. Paaß. From names to entities using thematic context distance. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, CIKM '11, page 857–866, New York, NY, USA, 2011. ACM.

[24] E. F. Tjong Kim Sang. Introduction to the CoNLL-2002 shared task: language-independent named entity recognition. In *proceedings of the 6th conference on Natural language learning - Volume 20*, COLING-02, page 1–4, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

[25] E. F. Tjong Kim Sang and F. De Meulder. Introduction to the CoNLL-2003 shared task: language-independent named entity recognition. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 - Volume 4*, CONLL '03, page 142–147, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

[26] D. Urbansky, J. A. Thom, D. Schuster, and A. Schill. Training a named entity recognizer on the web. In *Proceedings of the 12th international conference on Web information system engineering*, WISE'11, page 87–100, Berlin, Heidelberg, 2011. Springer-Verlag.

[27] C. Wang, K. Chakrabarti, T. Cheng, and S. Chaudhuri. Targeted disambiguation of ad-hoc, homogeneous sets of named entities. In *Proceedings of the 21st international conference on World Wide Web*, WWW '12, page 719–728, New York, NY, USA, 2012. ACM.

[28] H. J. Watson and B. H. Wixom. The current state of business intelligence. *Computer*, 40(9):96–99, 2007.

[29] A. Weichselbraun. A utility centered approach for evaluating and optimizing geo-tagging. In *First International Conference on Knowledge Discovery and Information Retrieval (KDIR 2009)*, pages 134–139, Madeira, Portugal, October 2009.

[30] A. Weichselbraun, S. Gindl, and A. Scharl. A context-dependent supervised learning approach to sentiment detection in large textual databases. *Journal of Information and Data Management*, 1(3):329–342, 2010.

[31] A. Weichselbraun, S. Gindl, and A. Scharl. Extracting and grounding context-aware sentiment lexicons. *IEEE Intelligent Systems*, 28(2):39–46, 2013.

[32] A. Weichselbraun, S. Gindl, and A. Scharl. Enriching semantic knowledge bases for opinion mining in big data applications. *Knowledge-Based Systems*, 2014. forthcoming (accepted 26 April 2014).

[33] F. Wu and D. S. Weld. Open information extraction using wikipedia. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, page 118–127, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[34] M. A. Yosef, S. Bauer, J. Hoffart, M. Spaniol, and G. Weikum. Hyena-live: Fine-grained online entity type classification from natural-language text. In *ACL (Conference System Demonstrations)*, pages 133–138. The Association for Computer Linguistics, 2013.