Linked Enterprise Data for Fine Grained Named Entity Linking and Web Intelligence

Albert Weichselbraun
Swiss Institute for Information Research
University of Applied Sciences
Chur, Switzerland
albert.weichselbraun@htwchur.ch

Daniel Streiff
Swiss Institute for Information Research
University of Applied Sciences
Chur, Switzerland
daniel.streiff@htwchur.ch

Arno Scharl
Department of New Media Technology
MODUL University Vienna
Vienna, Austria
scharl@modul.ac.at

ABSTRACT

To identify trends and assign metadata elements such as location and sentiment to the correct entities, Web intelligence applications require methods for linking named entities and revealing relations between organizations, persons and products. For this purpose we introduce Recognyze, a named entity linking component that uses background knowledge obtained from linked data repositories. This paper outlines the underlying methods, provides insights into the migration of proprietary knowledge sources to linked enterprise data, and discusses the lessons learned from adapting linked data for named entity linking. A large dataset obtained from Orell Füssli, the largest Swiss business information provider, serves as the main showcase. This dataset includes more than nine million triples on companies, their contact information, management, products and brands. We identify major challenges towards applying this data for named entity linking and conduct a comprehensive evaluation based on several news corpora to illustrate how Recognyze helps address them, and how it improves the performance of named entity linking components drawing upon linked data rather than machine learning techniques.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*Data Mining*; H.4.2 [**Information Systems Applications**]: Types of Systems—*Decision support*

General Terms

Algorithms, Design, Experimentation

Keywords

linked open data, linked enterprise data, named entity linking, business news, Web intelligence

.

1. INTRODUCTION

First coined by Hans Peter Luhn [13], the term *business intelligence* gained prominence in the 1990s when Howard Dresner from the Gartner Group started using it in its current interpretation [21]. Business intelligence is considered an umbrella term describing concepts and methods to improve business decision making by using fact-based support systems [4], which typically combine 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 [14].

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 customizing products to consumer needs. This potential motivated companies to apply Web intelligence methods to analyzing blogs, product reviews and social media streams. State-of-the-art Web intelligence systems deploy data mining engines for extracting structured data from such unstructured textual sources [3]. Named entity linking is a crucial task in this process, ensuring that the extracted information is assigned to the correct entities such as persons, organizations or products. To support this task, this paper introduces Recognyze as a novel component for named entity linking that draws upon background knowledge from structured sources (e.g. linked data) to achieve a high level of accuracy.

The paper is structured as follows: Section 2 discusses related work on named entity recognition and information extraction approaches that leverage background knowledge for optimizing their performance. Section 3 then describes the linked enterprise data repository underlying this paper and shows how Recognyze draws upon the knowledge contained in this repository. A detailed evaluation in Section 4 is followed by an outlook and conclusions in Section 5.

2. RELATED WORK

This section provides an overview of related work in named entity recognition and information extraction approaches that draw upon background knowledge retrieved from structured sources to improve their performance.

2.1 Named Entity Recognition

Named entity recognition identifies references to named entities in unstructured documents and classifies them into categories such as locations, persons and organizations. Urbansky et al. [19] 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 such as clustering.

Many approaches either use Wikipedia for training their models [12, 15] or draw upon background knowledge retrieved from Wikipedia to improve the accuracy of the named entity disambiguation process [9, 16, 10]. Han and Zhao [9] observe an improvement of 10.7% over traditional bag-of-word approaches, and a 16.7% improvement over traditional social network-based disambiguation methods.

Hoffart et al. [10] harness context information from structured data sources such as DBpedia and YAGO, and introduce 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 that occur in a document.

Pilz and Paaß [16] 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. [12] 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, exploit Wikipedia annotations, and uses Wikipedia's category hierarchy for capturing co-occurrence probabilities among entities.

Recently, Nothman et al. [15] used Wikipedia to create multi-lingual training data for named entity recognition tasks. Their approach yielded 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 [17, 18] shows that they outperform a number of other approaches to automatic named entity recognition.

Fernández et al. [6] present IdentityRank, a supervised algorithm for disambiguating names in news coverage. The authors leverage 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 [11] explores how named entity recognition methods can be applied to challenging datasets such as those derived from social media streams, which are characterized by short and often noisy text.

2.2 Named Entity Linking

Named entity linking, which is also known as named entity resolution, not only classifies named entities but also grounds them to a knowledge base such as DBpedia and Wikipedia, or to a relational database. Gangemi[7] provides an overview of knowledge extraction tools including specific applications for named entity linking. Wang et al. [20] 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 distinguish terms such as Apple or HP from their ambiguous counterparts when they occur in documents with an information technology or business focus.

2.3 Background Knowledge for Information Extraction

Hoffart et al. [10] and Weichselbraun et al. [24] demonstrate that considering external knowledge for information extraction tasks such as named entity recognition can significantly improve the accuracy of the deployed methods.

Opinion mining, a research field that automatically assesses text sentiment, extracts sentiment targets and aspects influencing the text's polarity. 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[1], SenticNet[2] and SentiWordNet[3]. Recent research in this area shows how methods that have been enhanced with the ability to draw upon background knowledge are able to (i) adapt their evaluations to the text's context [23, 5], (ii) distinguish between ambivalent concepts [24] 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 [25] 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 between 18% and 34% of the F-measure when compared to a similar approach that solely relied on hand crafted heuristics for generating training data.

3. METHOD

The Recognyze component introduced in this paper identifies named entities in unstructured documents of heterogeneous origin, and links these entities to structured sources. This section first describes the linked open data and linked enterprise data repositories used in this work, and then elaborates on how these repositories are leveraged in the disambiguation and named entity linking process.

3.1 Linked Open Data

Recognyze draws upon public and enterprise linked open data repositories for disambiguating named entities. We use 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 persons and organizations, and GeoNames[4] for recognizing geographic locations.

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

Table 1: Vocabulary used for the Orell Füssli Wirtschaftsinformationen linked enterprise data repository.

Namespace	number of elements	examples
dbprop	2	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

3.2 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, and reusing well-known vocabularies such as Friend-of-a-Friend (FOAF), Dublin Core (DC) and Simple Knowledge Organization System (SKOS).

The presented work draws upon such a linked enterprise data repository, created together with the Orell Füssli Business Information AG (OFWI), Switzerland's largest provider of business information. We have integrated data on more than 2.9 million companies with data sources on additional company names, the company's senior managers, product, address and contact information, business figures such as number of employees and turnover, brands offered by the company and the industry sector the company operates in. Removing duplicates and references to inactive companies yielded a linked enterprise data repository with more than 9 million triples.

To ensure interoperability with public resources, we have used well-known linked open data vocabularies wherever possible. Table 1 presents an overview of the used namespaces, the total number of elements taken from a particular namespace, and a number of selected example elements. The repository entirely draws upon 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.

3.3 Major Challenges

This section will discuss the major challenges and obstacles of using linked enterprise data for named entity linking. For this purpose, the following terminology 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 possible references to a named entity in text documents, often derived from legal company names.

3. *ambiguous search terms*, such as "Apple" are search needles that are considered ambiguous.

4. *unambiguous search terms* are considered specific enough to prevent ambiguities.

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.).

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 dataset includes very small companies. The German version of Wikipedia, for instance, lists three companies with the ambiguous name "Total" as of October 2013. 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 2 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.

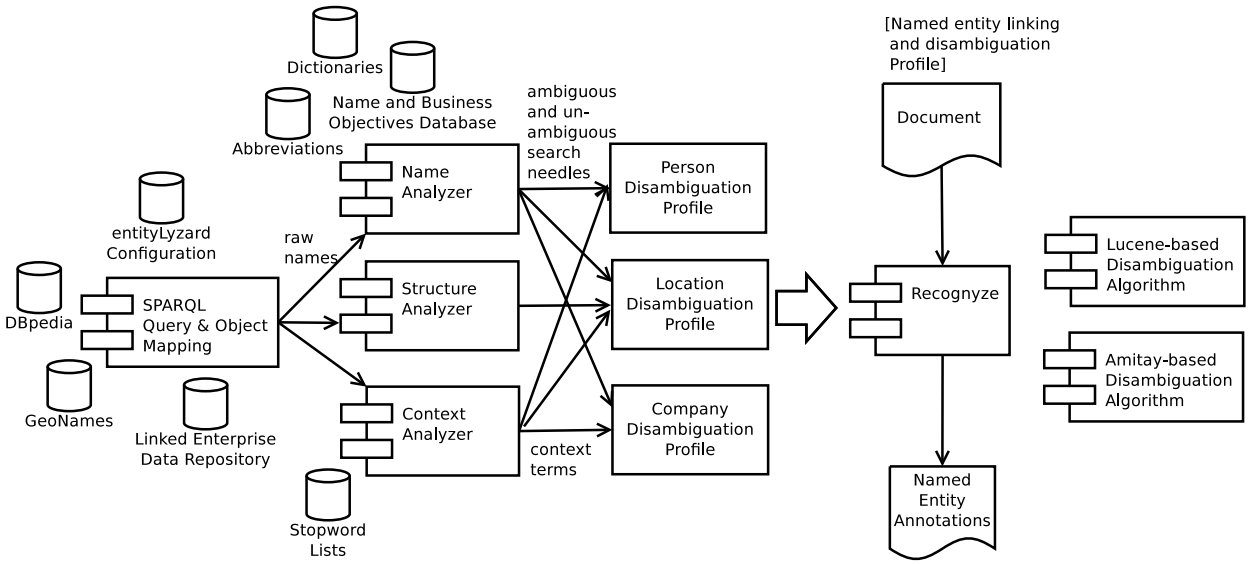

Figure 1: Named entity linking with Recognyze.

3.4 Named Entity Linking

Figure 1 shows how Recognyze draws upon 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 linked data 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 data repository as well as the mapping of these data onto the corresponding classes, pre-processors, and disambiguation algorithms.

3.4.1 Name Analyzer

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 documents, they rarely occur in documents relevant for Web Intelligence such as news articles, product forums and social media sites. Recognyze therefore includes a name analyzer component that decomposes legal company names into ambiguous and unambiguous search needles that resemble the most probable names used to refer to the company. This name analyzer uses an entropy-based heuristic to create search needles from the raw company name obtained from the knowledge source which are (i) short enough to occur in informal textual documents such as News articles, and (ii) unique enough to prevent ambiguities.

The component draws upon (i) a database of common Swiss, Italian, German and French last names, given names, business objectives, company types and abbreviations, (ii) heuristic rules which determine whether a word is most likely 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, and (iii) information on the number of different name component classes (e.g. abbreviation, name, dictionary term, trade) used in the search term. Our tests 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 Challenge 1.2. in Table 2). The name analyzer penalizes case insensitive names with negative initial entropy and by switching to case insensitive look-ups for the token classification. 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 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 the Swiss).

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 3 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.

3.4.2 Structure Analyzer

The *structure pre-processing* component is used to extract and integrate structural and hierarchical information into

Table 2: 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 the company names that are highly ambiguous. Example: Aktien (English: shares), Hell (English: bright), Maximum (English: maximum), etc.
1.2.	uppercase only company names	use of uppercase only company names; hard to find; complicates 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 on the raw data. 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), Aktien (English: shares), 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 uses 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 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.

3.4.3 Context Analyzer

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

3.4.4 Disambiguation and Ranking

The *Recognyze disambiguation process* draws upon the disambiguation profiles created by the processing of the knowledge base (Figure 1). Agents that call Recognyze have to specify the incoming documents and the named entity linking and disambiguation profile to be be applied for the named entity linking process. To identify candidate mentions 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 analyzer,
- 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.

Table 3: Automatic mappings of legal company names to search needles produced by the name analyzer.

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

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 organizations and persons.

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

Per entity queries q_e represent needle sets for an entity consisting of unambiguous company names, ambiguous company names, and context terms and their corresponding weights obtained from the pre-processing. The inverse document frequency ($idf(t)$) value ensures that rare terms provide a higher contribution to the total score and d refers to the document in which the needles have been found.

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 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 recognition 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 persons and organizations as well.

4. EVALUATION

The algorithms used for identifying locations have been thoroughly described in earlier work [22]. 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 persons 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).

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.

4.1 Data Sources

The evaluation has been performed on the following datasets:

1. The *AWP.ch business news* dataset 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 dataset* which consists of 150 randomly selected German-speaking news messages which have been manually annotated by domain experts. The annotations cover *all* companies in a particular document.

3. The *NZZ (Neue Zürcher Zeitung) news dataset* 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 datasets to contrast Recognyze's named entity linking performance for documents from rather formal business news (AWP dataset), as compared to documents from less formal newspaper articles (NZZ dataset). The latter cover a much larger range of topics and are, therefore, expected to be more prone to ambiguities.

4.2 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 Analyzer (Section 3.4.1) 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.

2. to which extend Recognyze considers context information, and

3. the strategy used for ranking articles, with or without re-scoring (Section 3.4.4). Recognyze uses context information for disambiguation and entity ranking. 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.

4.3 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 draws upon 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.

4.4 Results

Table 4 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 4: Recognyze named entity linking and ranking performance on the full AWP dataset.

name processing	context	R@1
Raw names	·	0.60
	✓	0.69
Simple name pre-processing	·	0.56
	✓	0.53
Filtering of ambiguous results	·	0.62
	✓	0.72
Name pre-processing	·	0.65
	✓	0.73
Name pre-processing & filtering	·	0.68
	✓	0.72

Table 5 illustrates the influence of the domain on the use of company names. The evaluation draws upon 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.

Applying the pre-processing techniques discussed in Section 3 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 5: 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

Table 6 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 draw upon 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 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 6 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.

4.5 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 name analyzer marked the company name as ambiguous and the text only contained the ambiguous name without any of the prefixes or 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) 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 4), two additional error source 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 datasets 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. [8] 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. [6] report a *disambiguation* accuracy of 96% for their named entity disambiguation approach. This accuracy has

Table 6: Recognyze named entity linking performance on the extended NZZ and AWP datasets.

name processing	context	rescore	AWP messages P	AWP messages R	AWP messages F1	NZZ articles P	NZZ articles R	NZZ articles 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 & filtering	·	·	0.15	0.62	0.19	0.07	0.50	0.11
	·	✓	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 & filtering	·	·	0.36	0.71	0.41	0.38	0.75	0.46
	·	✓	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

been measured for the disambiguation process (but not for the overall named entity recognition), 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 October 2013) or 28 companies with the name "Absolut" (versus three on 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).

5. OUTLOOK AND CONCLUSIONS

This paper presented Recognyze, a named entity linking component that draws on background knowledge from linked open data sources such as DBpedia and GeoNames, or from enterprise linked data. In contrast to other approaches, Recognyze does not apply machine learning and therefore does not require training corpora or iterative learning steps. An entropy-based name analyzer extracts relevant company names, context and structure analyzers obtain contextual and structural information which is then used for named entity linking and ranking.

The article discusses problems encountered when using external data sources, and presents methods for addressing them. The high recall for named entities referenced in business documents can be attributed to the use of a comprehensive linked enterprise repository containing detailed background knowledge to support the named entity recognition process. The recall is lower when processing more informal sources such as news articles, but can be improved through the pre-processing steps introduced in this paper.

Recognyze's overall named entity linking performance is quite respectable. Although the literature reports higher accuracies for named entity linking methods that apply machine learning techniques, Recognyze provides a suitable alternative to these approaches since, it

1. is not limited to a particular knowledge source,
2. does not require any training steps 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 a 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.

The evaluation thus supports the claim that Recognyze successfully disambiguates and grounds named entities in settings where a lot of similarly named alternatives (such as for instance the ambiguous company names Total, or Absolut) and collisions with common terms such as "sich bewusst sein" (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 named entity disambiguation process. We will also optimize and evaluate disambiguation profiles that work with publicly available linked open data sources such as DBpedia.

Acknowledgment

The research presented in this paper has been conducted as part of the COMET Project (www.htwchur.ch/comet), funded by the Swiss Commission for Technology and Innovation (KTI), 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.

6. 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] 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.

[6] 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.

[7] 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.

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

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] H. P. Luhn. A business intelligence system. *IBM Journal of Research and Development*, 2(4):314–319, 1958.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

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

[22] 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.

[23] 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.

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

[25] 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.