Context-Enhanced Adaptive Entity Linking

Filip Ilievski[∓], Giuseppe Rizzo[*], Marieke van Erp[∓], Julien Plu[°], Raphaël Troncy[°]
[∓]Vrije Universiteit Amsterdam, Netherlands {f.ilievski,marieke.van.erp}@vu.nl
[*]ISMB, Italy giuseppe.rizzo@ismb.it
[°]EURECOM, France {plu,troncy}@eurecom.fr

Abstract

More and more knowledge bases are publicly available as linked data. Since these knowledge bases contain structured descriptions of real-world entities, they can be exploited by entity linking systems that anchor entity mentions from text to the most relevant resources describing those entities. In this paper, we investigate adaptation of the entity linking task using contextual knowledge. The key intuition is that entity linking can be customized depending on the textual content, as well as on the application that would make use of the extracted information. We present an adaptive approach that relies on contextual knowledge from text to enhance the performance of ADEL, a hybrid linguistic and graph-based entity linking system. We evaluate our approach on a domain-specific corpus consisting of annotated WikiNews articles.

Keywords: Adaptive, Contextual, Entity Linking, Knowledge Extraction

1. Introduction

The Linked Open Data (LOD) cloud,[1] i.e. the union of numerous interlinked RDF datasets published following the linked data principles, represents a remarkable source of real-world knowledge. As such, it can be exploited by intelligent systems to extract information from textual resources. In particular, an intelligent system that anchors entities from text to existing entities in a knowledge base, namely an entity linking system, could benefit from the structured semantic knowledge describing those entities in the knowledge base. This knowledge enables an entity linking system to model and perform reasoning over the semantic context of the real-world entities it links to.

However, the heterogeneity of the LOD knowledge also represents a barrier towards its automatic use, integration and manipulation. Entity linking systems have thus difficulties to adapt across domains, i.e. to perform unified processing of textual resources from different domains. This has inspired the development of domain-focused systems that perform reasonably well on concrete, domain-dependent information extraction tasks, but demonstrate low accuracy when applied in different domains.

The key intuition we follow in this paper is that entity linking can be customized for new domains by making use of contextual knowledge associated with the textual content to analyze and the task that will make use of the extracted information. We investigate how to enhance the cross-domain performance of entity linking through **contextual adaptation**, i.e. the integration of semantic knowledge from a domain-specific knowledge base with evidence from additional information sources (the text to analyze and the task to address).

Our approach aims to enhance the cross-domain performance of ADEL, a hybrid linguistic and graph-based entity linking module by applying a set of heuristics for contextual adaptation. These heuristics consider the order of processing of text, coreference of entity mentions, topical domain relevance and semantic typing. We assess the impact of our approach on an automotive domain dataset, the MEANTIME corpus (Minard et al., 2016), as well as on the standard AIDA-YAGO2 dataset (Hoffart et al., 2011).

The remainder of this paper is structured as follows. We discuss the role of adaptation in information extraction tasks in Section 2.. We describe our approach in Section 3.. In Section 4., we provide statistics about the benchmark datasets used in the experimental settings, while in Section 5., we report on the obtained results. We analyze these results in Section 6.. An overview of the strengths and weaknesses evidenced by our results so far, along with future work, are reported in Section 7.. To facilitate reuse and replicability, we provide our code and resources at: `https://github.com/MultimediaSemantics/relink`.

2. Related Work

Exploitation of contextual knowledge is essential in order to compensate the degradation in performance of an Information Extraction (IE) approach when it processes data that follows a different distribution than the one used for training a system. In (Heuss et al., 2014), the authors demonstrate that the topical domain specification has a significant impact on the Named Entity Recognition (NER) performance, leading to drops in terms of F_1 measures up to 60% $\pm$ 30% when compared to common extraction scenarios. These results indicate that it is typically preferable to use in-domain data.

Ongoing research efforts in the IE field have been investigating the effects of using contextual knowledge associated with textual data. (Rizzo et al., 2014) propose a supervised learning approach, trained with labeled data which represents different textual genres. The main idea is to exploit supervised learning systems trained on different data in order to improve an existing domain-specific entity recognition and linking system, by boosting its recall and widening its content coverage. Other attempts, such as (Wu et al., 2009), use contextual knowledge adaptation to improve the accuracy of a named entity classifier in recognizing entities that do not belong to the distribution of the labeled

[1] `http://lod-cloud.net`

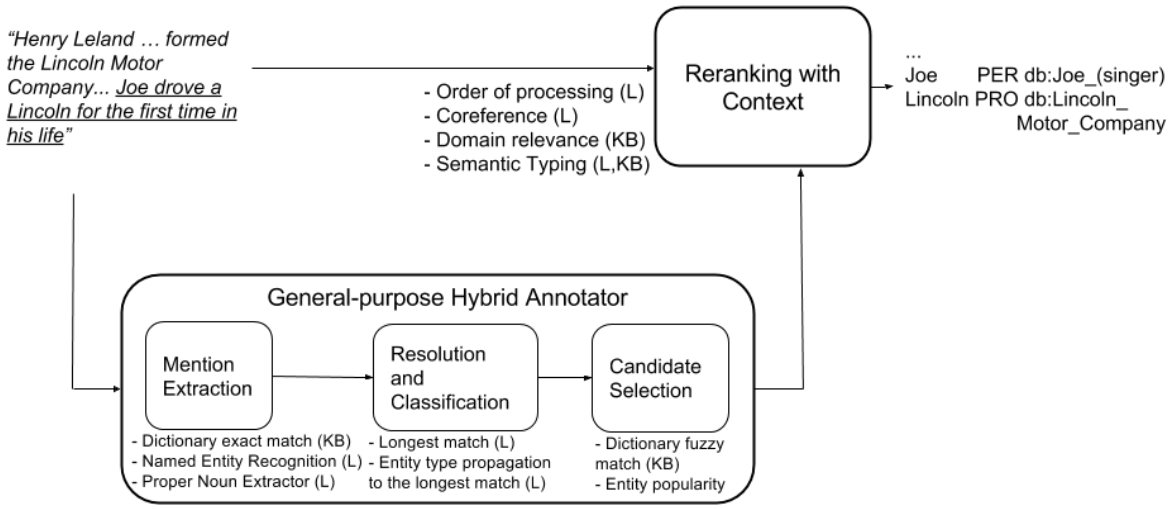

Figure 1: Workflow of the approach of adaptive entity linking exploiting contextual knowledge.

data used in the learning stage. This approach follows rigid designators, where a small (and therefore incomplete) set of labeled data is provided, annotated by a small set of annotators.

Contextual adaptation has been proposed in recent entity linking tasks. Among the first of these approaches, the authors in (Zhang et al., 2015) select entities that are related to the document topics using the relation between the candidate entities and their associated categories. Our method addresses the task differently, relying on *a priori* set of topical and genre classes, that are used to further refine the selection of the referent entity in the knowledge base.

On the other hand, an important gap concerning the generalizability of entity linking systems over knowledge bases has been spotted by (Usbeck et al., 2014). They propose AGDISTIS, a system which performs knowledge base-agnostic disambiguation of named entities. The key rationale is that in order to develop Web-wide entity linking solutions, it is crucial to ensure these can easily be adapted to anchor with respect to new knowledge bases. As a positive side effect, this direction decreases the dependency of the developed systems on a single (or a small amount of) popular knowledge base(s), such as DBpedia. While our approach shares this goal to develop generally applicable entity linking systems, we also require that such generic system provides easy means for adaptation in various concrete domains.

3. Adaptive Entity Linking Approach

We propose an adaptive entity linking approach consisting of two steps (Figure 1): a general-purpose hybrid module and a domain adaptation module. We describe both steps in this Section, and depict how they operate on an example text.

3.1. General-purpose hybrid annotator

In the first step, we use a hybrid module that combines linguistic and graph-based algorithms to detect entity mentions in text and link them to existing DBpedia resources (Plu et al., 2015). This hybrid annotator consists of three main stages: *i)* Mention Extraction, *ii)* Resolution and Classification, and *iii)* Candidate Selection.

3.1.1. Mention Extraction

This stage detects mentions in text that are likely to denote entities. It is composed of two different modules: *extractors* and *overlap resolution*. The extractors module makes use of the following components: Part-of-Speech (POS) tagging nouns, and Named Entity Recognition (NER) classifying entities. In details, we use the Stanford NLP POS-Tagger (Toutanova et al., 2003) with the *english-bidirectional-distsim* model. We use the Stanford NER Tagger (Finkel et al., 2005) using the *NERClassifierCombiner* functionality to combine multiple CRF models together.

3.1.2. Resolution and Classification

The parallel strategy used in the mention extraction may generate overlaps. We make use of an *overlap resolution* module that takes the output of each component of the extractors module and decides on a single output with no overlaps. For instance, given two overlapping mentions *States of America* from Stanford NER and *United States* from Stanford POS tagger, we perform a union operation over the two phrases to obtain the mention *United States of America*. The type of the mention is then assigned according to the type assigned on the match or partial match by the entity recognizer.

3.1.3. Candidate Selection

This stage is composed of two sub-tasks, namely entity linking and candidate ranking. For the entity linking, we perform a lookup for an entity mention in an index built on top of both DBpedia2015-04[2] and a dump of the Wikipedia articles from February 2015[3] to generate potential candidates for an entity mention. This process can generate more than one link for mention. We then rank the links according

[2] http://wiki.dbpedia.org/services-resources/datasets/datasets2015-04

[3] https://dumps.wikimedia.org/enwiki

to Equation 1 and we consider the first ranked link to be the most suitable resource to disambiguate the entity.

The ranking function *r(l)* combines: *i)* the Levenshtein distance L between the entity mention m and a knowledge base label (e.g. Wikipedia title), *ii)* the maximum Levenshtein distance between the mention m and a label (title) of every Wikipedia redirect page from a set R, *iii)* the maximum distance between the mention m and every label (title) in the set of Wikipedia disambiguation pages D, *iv)* the PageRank (Page et al., 1999) PR value for every entity candidate l.[4] The weights a, b and c satisfy the following equation: $a + b + c = 1$ (as a convex combination) and $a > b > c > 0$. We empirically assessed that the string distance measure between a mention and a title is more important than the distance measure with a redirect page, which is itself more important than the distance measure with a disambiguation page.

As it is customary, we link the entities with no detected entry in a knowledge base to *NIL*. In cases when two or more candidate links attached to a single entity mention share the same maximum ranking score, we still rely on taking the first candidate in the ordered ranking list as the most probable candidate.

$$r(l) = (a \cdot L(m, title) + b \cdot max(L(m, R)) + c \cdot max(L(m, D))) \cdot PR(l) \quad (1)$$

3.2. Reranking with Context

In the second step, we apply the ReCon module,[5] which acts as a contextual adaptation layer over the hybrid module (Section 3.1.). ReCon adapts the entity linking task to the textual content that is being analyzed, by leveraging genre and topic domain information about the text. In the current version, ReCon focuses on entity linking in news articles.

ReCon uses the entity linking annotations performed by ADEL as a starting point and applies a set of heuristics which consider aspects of news articles that are not taken into account by the hybrid module. ReCon makes a reasonable assumption that the order of expressing information in news articles is somewhat systematic, i.e. the introductory sentences of an article tend to be written more explicitly, while the latter ones often contain abbreviations and incomplete mentions which are coreferential with previous explicit mentions. Furthermore, the title of an article is often intentionally ambiguous and uses figurative speech, thus requiring a reader to understand the running text of the article first. Finally, the entities mentioned in domain-specific news articles usually stem from that domain, e.g. in the basketball domain, it is customary to refer to the Chicago Bulls' mega-star Michael Jordan with "Jordan".

ReCon applies four article-wide heuristics which consider these properties of the newswire text and, additionally, allows them to be tunable to the topic domain of an article. Our heuristics take the form of binary rules. We detail them as follows:

H1: Order of processing. The first Recon heuristic, H1, focuses on the relation between the title and the running text of a news article. Concretely, we observe that the title of a news article is often ambiguous and uses figurative speech, probably because of the intention of the writer to make the title attractive for a generic reader. On the other hand, the running text of a news article, especially the introductory sentences, are written in a clear, informative and contextually rich manner, allowing a shallowly informed reader to grasp as much as possible from the presented news. As a reaction to these observations, we introduce our heuristic H1 which concerns the order of information extraction in a news article. Following H1, we first process the running text of a news article and disambiguate the title of the article at the end.

H2: Co-reference. Our second heuristic, H2, follows a related rationale. We observe that surface forms that refer to an entity earlier in text are surrounded by richer verbal context, compared to the latter mentions of the same entity. This is understandable from a practical perspective: once the writer has clearly introduced an entity, he can use abbreviations or more ambiguous ways to refer to it further in the text. As a consequence, our second heuristic also operates on a cross-sentence level and takes advantage of the order in which sentences and entities appear in text. Concretely, ReCon's H2 tries to detect earlier mentions of the same entity, i.e. H2 checks if an entity mention is in a same co-referential chain as another entity mention which occurred earlier in the same article. This co-reference relation is established based on the textual similarity of these mentions, in particular it is based on the following three rules:

1. An entity mention M_j is co-referential with a previously occurred mention M_i if M_j is an abbreviation of M_i
2. An entity mention M_j is co-referential with a previously occurred mention M_i if M_j is identical to M_i
3. An entity mention M_j is co-referential with a previously occurred mention M_i if M_j is a sub-string of M_i and M_i was linked to an entity of type `Person`

H3: Domain relevance. If no pre-occurring co-referential entity is found by H2, we apply a topic modeling heuristic, H3, in which we exploit a contextual knowledge base about the topic of interest. We rely on this knowledge base to examine whether a mention has been frequently and dominantly associated with a certain entity disambiguation link within the specific topical domain. The frequency and dominance for each surface form are expressed through threshold values: a mention is resolved to an entity link according to H3 only if the mention appears sufficiently (with

[4] The PageRank scores for every DBpedia resource originate from (Reddy et al., 2014).

[5] ReCon stands for Reranking with Context

	Number of Articles	Number of Tokens	Number of Entities	Number of Links	Number of NILs	Number of Entity Types
airbus	30	3,620	614	414	200	5
apple	30	3,452	812	525	287	5
gm	30	3,641	760	526	234	5
stock	30	3,362	449	331	118	4
aida-yago2	231	46,435	5,616	4,485	1,131	4

Table 1: Statistics on no. of articles, no. of tokens, no. of entities, no. of links, no. of NILs (entities that do not have a referent in the knowledge base), and no. of types of the four sub-datasets of the MEANTIME and AIDA-YAGO2 testb corpora.

frequency above the threshold frequency) in the domain knowledge base, and a certain entity link is dominantly associated with that mention (the frequency distribution for that entity needs to be above the dominance threshold).

H4: Semantic typing. Whenever either of H2 or H3 proposes an entity disambiguation link for a mention, we apply a fourth heuristic, H4, to check whether its semantic type [6] corresponds to the textual context of that surface form. In practice, this is done by matching its semantic type from DBpedia against the entity type as specified by the hybrid entity linker in the first step.[7] If this comparison is successful and the proposed entity corresponds to the type constraint set by the hybrid module, then the mention is linked to the proposed entity. Otherwise, we decline the proposed entity link. When none of the ReCon heuristics is able to resolve an entity mention, then we refrain to the linking provided in the first, hybrid step.

3.3. Running example

We illustrate our approach on a news article snippet (Figure 2). This article stems from the MEANTIME dataset.[8]

The entity mentions **General Motors**$_2$ and **General Motors Corporation (GM)**$_3$ are correctly linked to *db:General_Motors*[9] by the hybrid approach (step 1). This is pressumably because these two surface forms refer to *db:General_Motors* using a phrase which is both customary and extensive. **General Motors Corporation (GM)**$_3$ even specifically introduces an abbreviation "GM", allowing a human reader to understand the further mentions of this abbreviation in the remainder of the article.

However, the first step of our approach is not able to correctly disambiguate the other 4 entity mentions from this snippet. This is partially due to the fact that this module does not take into account the global, article-wide context, but instead relies on a combination of string distance and the PageRank algorithm. Therefore, the error can be explained by the linking formula in Section 3.1.: e.g., the string distance score over the title, the redirect and the disambiguation pages between the entity mention *GM* and the entity candidate *db:Germany* (0.32879817) is higher than with the entity candidate *db:General_Motors* (0.21995464). The linking of these entities is improved in the second step by ReCon, which "memorizes" information which have been previously stated in the article and uses this to process the remainder of the article, in a human-like manner. Once the abbreviation for General Motors has been introduced, ReCon assumes that further mentions of this abbreviation are co-referential to its earliest mention in the running text. This combination of the heuristics H1 and H2 is able to improve the annotations of **GM**$_1$, **GM**$_4$ and **GM**$_6$, which have previously been disambiguated to *db:Germany* according to the scoring function of our step 1.

In this snippet, we also demonstrate the power of our domain relevance heuristic, H3. Following this heuristic, we find in a domain knowledge base that the phrase "US$" has been dominantly and frequently associated with *db:United_States_dollar* in the automative industry domain. This allows ReCon to rerank the decision of the hybrid approach and disambiguate **US$**$_5$ as *db:United_States_dollar*, instead of as *db:United_States*.

The four reranking decisions we describe are confirmed by the semantic typing heuristic of ReCon (H4). The semantic types of *db:General_Motors* and *db:United_States_dollar* correspond to the semantic types suggested by the hybrid module based on the textual context of these mentions, which are Organization and Miscellaneous, respectively.

4. Datasets

We test our approach on two different datasets, MEANTIME and AIDA-YAGO2. Table 1 provides general statistics of the two datasets such as the number of articles, tokens, entities, links, NIL entities and the different entity types.

4.1. MEANTIME

The MEANTIME corpus (Minard et al., 2016) was developed within the NewsReader project.[10] The corpus consists of 120 English WikiNews articles as well as their translations in Spanish, Italian and Dutch, all annotated with entities, events, temporal expressions and semantic roles. In the entity annotation layer, links to DBpedia resources are included. As the focus of the NewsReader project lies on the financial-economic domain, the chosen articles reflect this topic. The MEANTIME corpus articles are split into

[6] The `rdf:type` of the entity

[7] The incompleteness of semantic typing in our knowledge bases poses a challenge to H3. In the implementation of this heuristic, we assume that entities which do not belong to any of the main classes fit the type constraint posed by the hybrid step.

[8] We describe this dataset in detail in Section 4.

[9] PREFIX db: $< http://dbpedia.org/resource/ >$

[10] `http://www.newsreader-project.eu/results/data`

GM$_1$ posts first annual loss since 1992
January 27 , 2006
General Motors$_2$ logo .
General Motors Corporation (GM)$_3$ has posted its first annual loss since 1992 . **GM**$_4$ reported losing **US$**$_5$ 4.8 billion in the fourth quarter of 2005 and a total loss of $ 8.6 billion for the entire year . **GM**$_6$ admitted Thursday night that the loss could swell further as it pays pensions and healthcare costs to thousands of former workers .
...

Figure 2: Snippet from article number 31965 in MEANTIME's GM subcorpus

four sub-corpora revolving around the following core topics: *i)* Airbus Boeing, *ii)* Apple Inc., *iii)* General Motors, Ford and Chrysler, and *iv)* the Stock market.
The corpus was created according to the NewsReader guidelines for annotation at document level (Tonelli et al., 2014). Since annotation of text with information for a multitude of NLP tasks is extremely costly, only the first six sentences of each article are fully annotated. Therefore, we focus our evaluation on a comparative results of the performance of those sentences only. There are five entity types annotated in total in the MEANTIME corpora: PRO (product), LOC (location), FIN (financial), ORG (organization) and PER (person).

4.2. AIDA-YAGO2

The AIDA-YAGO2 dataset is built on top of the most prominent dataset in named entity recognition, namely the benchmark dataset that was created for the CoNLL-2003 Language-Independent Named Entity Recognition shared task (Tjong Kim Sang and Meulder, 2003). For this task, English and German news articles were annotated with named entities and made available to the research community to train and test their systems. The English training data consists of 946 articles, containing $203,621$ tokens. The test data consists of 231 articles, containing $46,435$ tokens. The data was annotated manually with named entities of types: person, location, organization and miscellaneous. Part-of-speech and chunk tags were added automatically. There is fairly little overlap of named entities between the training and test datasets: only 2.9% of the named entities that occur in the training data also occur in the test data. (Hoffart et al., 2011) annotated each entity mention in the CoNLL-2003 data set with links to YAGO and Wikipedia. For evaluation of our approach, we focus on the 231 articles in the English testb-portion of the corpus.

5. Experimental Results

We compute the accuracy of our hybrid module (Section 3.1.) with two different settings, one for each dataset. For the MEANTIME dataset the extraction component uses the POS and the NER extractors whereas for the AIDA-YAGO2 dataset it uses only the NER extractor (Section 3.1.). Next, we use the hybrid module as a baseline to measure the impact of various ReCon heuristics (Section 3.2.) on the entity linking accuracy.
We report four versions of this comparison (Tables 2, 3, 4 and 5), which differ in terms of matching strictness and inclusion of NIL entities in the evaluation. When it comes to matching, we report results for both exact and partial/fuzzy matching of entity mentions.[11] We report results that both include and exclude NIL entities. For all four comparisons, we present the accuracy of our approach in terms of precision, recall and F_1-measure. The last row in each of these four tables presents the performance of the overall two-step system, consisting of the hybrid module and all four ReCon heuristics, applied consecutively. For consistency, we use the evaluation scripts developed within the NewsReader project.[12]

	Airbus	Apple	GM	Stock	AIDA-YAGO2
H1+H2	38	41	48	18	258
H1+H3	353	235	357	252	2972
H1+H2+H3	353	249	358	254	3045
H1+H2+H3+H4	297	236	313	233	2461

Table 6: Number of re-assigned links by various ReCon heuristics combinations

For each corpus, we also note the number of surface forms whose linked entity has been re-assigned based on the heuristics of ReCon in Table 6. As we discuss in Section 3.2., ReCon does not rerank for every surface form, but only for those surface forms that satisfy one of its heuristics. Because of this, the number of rerankings on a corpus is lower than the overall number of entity surface forms in that corpus. Additionally, the reranking decision by ReCon may coincide with the disambiguation of the hybrid step, which means that not every reranking decision performs alteration of the disambiguation link assigned by the hybrid module.

6. Results Analysis

The results reported in the Section 5. provide evidence that the task of entity linking can benefit from domain-specific knowledge. For most corpora, the domain-aware heuristics of ReCon are able to enhance the entity linking and improve its accuracy: the heuristics of ReCon bring improvement over the baseline module for 4 out of 5 corpora, irrespective of the evaluation settings (matching type or inclusion of NIL entities).

[11] In the fuzzy matching mode, we consider an entity mention as matched if at least one of the tokens overlaps between an entity mention from the system output and the gold standard. This type of matching thus tolerates partially incorrect entity boundaries.

[12] `https://github.com/newsreader/evaluation/tree/master/ned-evaluation`

	Airbus			Apple			GM			Stock			AIDA-YAGO2		
	p	r	F_1	p	r	F_1	p	r	F_1	p	r	F_1	p	r	F_1
Hybrid	38.81	26.81	31.71	10.07	5.14	6.81	30.36	16.16	21.09	28.73	15.71	20.31	46.82	41.36	43.92
Hybrid+H1+H2	38.81	26.81	31.71	10.07	5.14	6.81	33.21	17.68	23.08	28.73	15.71	20.31	46.35	40.96	43.49
Hybrid+H1+H3	43.06	29.95	35.33	10.04	5.14	6.8	40.62	22.24	28.75	27.81	15.71	20.08	55.58	49.97	52.62
Hybrid+H1+H2+H3	43.06	29.95	35.33	10.04	5.14	6.8	43.75	23.95	30.96	29.95	16.92	21.62	55.34	49.77	52.41
Hybrid+ReCon	42.36	29.47	34.76	10.04	5.14	6.8	38.95	21.1	27.37	25.13	14.2	18.15	52.93	47.56	50.1

Table 2: Exact mention match evaluation, excluding NIL entities. H1 (order of processing), H2 (co-reference), H3 (domain relevance) and H4 (semantic typing) correspond to the ReCon components described in Section 3.2.. Figures are in percentage.

	Airbus			Apple			GM			Stock			AIDA-YAGO2		
	p	r	F_1	p	r	F_1	p	r	F_1	p	r	F_1	p	r	F_1
Hybrid	58.74	40.58	48	19.78	10.09	13.37	50.36	26.81	34.99	59.12	32.33	41.8	49.14	43.41	46.1
Hybrid+H1+H2	59.09	40.82	48.29	19.78	10.09	13.37	55	29.28	38.21	59.12	32.33	41.8	48.67	43.01	45.67
Hybrid+H1+H3	62.5	43.48	51.28	20.07	10.29	13.6	63.54	34.79	44.96	66.31	37.46	47.88	57.89	52.04	54.81
Hybrid+H1+H2+H3	62.15	43.24	51	20.07	10.29	13.6	67.36	36.88	47.67	68.98	38.97	49.81	57.65	51.84	54.59
Hybrid+ReCon	61.46	42.75	50.43	20.07	10.29	13.6	62.1	33.65	43.65	63.1	35.65	45.56	55.21	49.61	52.26

Table 3: Fuzzy mention match evaluation, excluding NIL entities. H1 (order of processing), H2 (co-reference), H3 (domain relevance) and H4 (semantic typing) correspond to the ReCon components described in Section 3.2.. Figures are in percentage.

	Airbus			Apple			GM			Stock			AIDA-YAGO2		
	p	r	F_1	p	r	F_1	p	r	F_1	p	r	F_1	p	r	F_1
Hybrid	36.22	18.4	24.41	8.88	3.32	4.84	27.83	11.32	16.09	26.63	11.8	16.36	44.4	40.15	42.17
Hybrid+H1+H2	36.22	18.4	24.41	8.88	3.32	4.84	30.42	12.37	17.59	26.63	11.8	16.36	44.02	39.81	41.81
Hybrid+H1+H3	40.38	20.52	27.21	8.88	3.32	4.84	38.19	15.53	22.08	26.63	11.8	16.36	52.00	47.03	49.39
Hybrid+H1+H2+H3	40.38	20.52	27.21	8.88	3.32	4.84	41.1	16.71	23.76	28.64	12.69	17.59	51.80	46.85	49.20
Hybrid+ReCon	39.74	20.19	26.78	8.88	3.32	4.84	36.25	14.74	20.95	24.12	10.69	14.81	49.85	45.08	47.35

Table 4: Exact mention match evaluation, including NIL entities. H1 (order of processing), H2 (co-reference), H3 (domain relevance) and H4 (semantic typing) correspond to the ReCon components described in Section 3.2.. Figures are in percentage.

	Airbus			Apple			GM			Stock			AIDA-YAGO2		
	p	r	F_1	p	r	F_1	p	r	F_1	p	r	F_1	p	r	F_1
Hybrid	56.41	28.67	38.01	18.09	6.77	9.86	47.57	19.34	27.5	56.28	24.94	34.57	48.18	43.57	45.76
Hybrid+H1+H2	56.73	28.83	38.23	18.09	6.77	9.86	51.78	21.05	29.93	56.28	24.94	34.57	47.80	43.23	45.40
Hybrid+H1+H3	60.26	30.62	40.6	18.42	6.9	10.04	60.19	24.47	34.8	63.82	28.28	39.2	55.78	50.44	52.98
Hybrid+H1+H2+H3	59.94	30.46	40.39	18.42	6.9	10.04	63.75	25.92	36.86	66.33	29.4	40.74	55.58	50.27	52.79
Hybrid+ReCon	59.29	30.13	39.96	18.42	6.9	10.04	58.58	23.82	33.86	60.8	26.95	37.35	53.63	48.50	50.94

Table 5: Fuzzy mention match evaluation, including NIL entities. H1 (order of processing), H2 (co-reference), H3 (domain relevance) and H4 (semantic typing) correspond to the ReCon components described in Section 3.2.. Figures are in percentage.

Looking at specific corpora, we observe the highest improvement on the GM corpus, while the domain adaptation of ReCon seems to be not effective on the Apple corpus. For these two corpora, there is no correlation between the number of re-assigned entity links by ReCon and the corresponding F1-measures. An intuitive interpretation of the discrepancy in the accuracy between these corpora would be that GM stems from the ideal domain that is intended to be processed by ReCon, because the topics of GM revolve around the automotive industry which is exactly the topic of the knowledge used in ReCon's H3.

The topics covered by AIDA-YAGO2 do not stem from the automotive industry, thus ReCon's improvement there may seem counterintuitive. However, AIDA-YAGO2 contains many well-known entities from a neutral domain (e.g. *db:France*), some of which tend to be very frequent and dominant also in the automotive domain. This causes such entities to be resolved correctly by H3. From an entity overlap analysis of the corpora, we also know that 23.30% of the MEANTIME entities appear in AIDA-YAGO2.

Our results indicate that, in most cases, the highest score is achieved by combining heuristics H1 and H3, or by combining H1, H2 and H3. This leads to several conclusions. Firstly, we observe that H3 is the most successful individual heuristic. This heuristic brings a key complementary information to the first, hybrid step: while the hybrid step relies on a notion of popularity through the PageRank algorithm, H3 indicates how often and how dominantly a cer-

tain surface form is used to refer to a certain entity within a concrete domain (in this paper, we compute these statistics on over a million domain-specific news articles). Next to its potential, this heuristic is also applied more frequently than the other heuristics, for instance, H3 is triggered up to 10 times more often than H2 on the corpora we analyze (Table 6). Secondly, the order of processing (H1), i.e. starting with the clearest and most explicitly conveyed information, and then proceeding towards increasingly ambiguous sentences, helps the entity linking system similarly as it helps humans. While we do not quantify the effect solely of this heuristic, qualitative case-by-case analysis such as the one in Section 3.3., provides evidence for the usefulness of this heuristic. In the case of the GM corpus, this heuristic helps the co-reference heuristic (H2) reach a higher score. Thirdly, we note that the semantic typing heuristic worsens the results. This is unfortunate, as we do think that semantic typing adds an interesting perspective to the entity linking process. In practice, however, the semantic typing heuristic is sensitive to two key factors which directly influence its performance: *i)* accuracy of the semantic typing of the entity with respect to its textual context, which is automatically performed by the hybrid module; *ii)* accuracy of the entity types in DBpedia (if any).[13] Hence, an optimal application of the semantic typing heuristics should be examined further in future work.

Furthermore, we observe a single pattern for the accuracy of our heuristics per corpus which is confirmed across Tables 2-5, irrespectfully of the recognition strictness or the inclusion of NIL entities. Namely, if a certain combination of heuristics improves the entity linking accuracy on a certain corpus, this is manifested in each of the evaluation results tables. It is also not surprising that the accuracy of linking on partially matched entities is better than the accuracy on exactly matched entities. The choice of whether to include NIL entities seems to be of different importance across datasets. For the MEANTIME corpora, this is a very important evaluation decision, while in AIDA-YAGO2 the NIL entities have much less influence. This can be explained by the results in Table 1, where we observe that between 26.28% and 35.34% of the entity mentions in the MEANTIME dataset are annotated with a NIL link. The percentage of NIL entities in AIDA is lower (20.14%).

7. Conclusions and Future Work

In this paper, we investigate the role of contextual knowledge as a domain adaptation layer for the task of entity linking. Our hypothesis is that the task of entity linking can be customized by taking into account genre and topical domain information about the text to analyze. We propose an adaptive context-aware solution, consisting of two steps and a set of rule-based heuristics. Our system focuses on entity linking in news articles and it can be tuned to a topic through leveraging a topic-specific knowledge source. We provide evidence for the potential of our approach on the MEANTIME dataset, which was created in order to evaluate entity linking in the automotive industry domain, as well as on the well-known AIDA-YAGO2 dataset.

We have demonstrated the promise of domain adaptation on the analyzed datasets through a qualitative analysis in Section 3.3. and the quantitative analysis in Section 6. Still, our contextual reranker can easily benefit from an extended set of heuristics, which is expected to increase the importance of the interaction between the different heuristics and require more advanced decision making models than the current rule-based decision model. These heuristics, for instance, could help modelling the genre domain further, by making use of other properties of news articles.

We plan to proceed with the experiments to further exploit the topical domain of an article. While we already take the topical context into account to some extent (ReCon's H3), this can be considered further by leveraging contextual knowledge more extensively. However, we observe that for certain datasets it is not trivial to decide on the document topic. For instance, in Section 6. we have discussed that our heuristics, including H3, improve the accuracy both on the GM corpus and on AIDA-YAGO2, although the topic context of GM is by far more appropriate for our approach. We calculate that almost a quarter of the entities which occur in the MEANTIME corpus are also represented in AIDA-YAGO2, which provides evidence that the subset of well-known entities is prominent in datasets from various domains.

Finally, although the current version of ReCon supports customizability in terms of the knowledge source used, we have not experimented with different knowledge bases and choosing an appropriate knowledge source dynamically to match the topical domain of the analyzed news article. Investigation of this dynamic customizability of our approach lies in the realm of future work.

Acknowledgments

This work was partially supported by the European Union's 7th Framework Programme via the NewsReader Project (ICT-316404), by the European Union's H2020 Framework Programme via the FREME Project (644771), the innovation activity 3cixty (14523) of EIT Digital (`https://www.eitdigital.eu`) and the CLARIAH-CORE project financed by NWO (`http://www.clariah.nl`).

Bibliographical References

Finkel, J. R., Grenager, T., and Manning, C. (2005). Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*.

Heuss, T., Humm, B., Henninger, C., and Rippl, T. (2014). A Comparison of NER Tools W.R.T. A Domain-specific Vocabulary. In *10^{th} International Conference on Semantic Systems*, SEM.

Hoffart, J., Yosef, M. A., Bordino, I., Fürstenau, H., Pinkal, M., Spaniol, M., Taneva, B., Thater, S., and Weikum, G. (2011). Robust Disambiguation of Named Entities in Text. In *Empirical Methods in Natural Language Processing (EMNLP'11)*.

[13] (Paulheim and Bizer, 2014) report that large portion of the DBpedia entities are incomplete in terms of their semantic types and estimate 2.7 million missing type statements in DBpedia.

Minard, A.-L., Speranza, M., Urizar, R., Altuna, B., van Erp, M., Schoen, A., and van Son, C. (2016). MEANTIME, the newsreader multilingual event and time corpus. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC). TO APPEAR (https://goo.gl/ajyuRo)*.

Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). The PageRank citation ranking: bringing order to the web.

Paulheim, H. and Bizer, C. (2014). Improving the Quality of Linked Data Using Statistical Distributions. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 10(2):63–86.

Plu, J., Rizzo, G., and Troncy, R. (2015). Revealing Entities from Textual Documents Using a Hybrid Approach. In *3rd International Workshop on NLP & DBpedia*, NLP&DBpedia.

Reddy, D., Knuth, M., and Sack, H. (2014). Dbpedia graphmeasures. dataset. Publication date July 2014.

Rizzo, G., van Erp, M., and Troncy, R. (2014). Benchmarking the Extraction and Disambiguation of Named Entities on the Semantic Web. In *9th International Conference on Language Resources and Evaluation*.

Tjong Kim Sang, E. F. and Meulder, F. D. (2003). Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In *17th Conference on Computational Natural Language Learning (CoNLL'03)*, Edmonton, Canada.

Tonelli, S., Sprugnoli, R., and Speranza, M. (2014). NewsReader Guidelines for Annotation at Document Level, version 4.1. Technical report, Fondazione Bruno Kessler.

Toutanova, K., Klein, D., Manning, C. D., and Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*.

Usbeck, R., Ngomo, A.-C. N., Röder, M., Gerber, D., Coelho, S. A., Auer, S., and Both, A. (2014). Agdistis-graph-based disambiguation of named entities using linked data. In *The Semantic Web–ISWC 2014*, pages 457–471. Springer.

Wu, D., Lee, W. S., Ye, N., and Chieu, H. L. (2009). Domain Adaptive Bootstrapping for Named Entity Recognition. In *Conference on Empirical Methods in Natural Language Processing*, EMNLP.

Zhang, L., Liu, C., and Rettinger, A. (2015). A Topic-Sensitive Model for Salient Entity Linking. In *3rd International Workshop on Linked Data for Information Extraction*, LD4IE.