Entity Typing using Distributional Semantics and DBpedia

Marieke van Erp and Piek Vossen

Vrije Universiteit Amsterdam
{marieke.van.erp,piek.vossen}@vu.nl

Abstract. Recognising entities in a text and linking them to an external resource is a vital step in creating a structured resource (e.g. a knowledge base) from text. This allows semantic querying over a dataset, for example selecting all politicians or football players. However, traditional named entity recognition systems only distinguish a limited number of entity types (such as Person, Organisation and Location) and entity linking has the limitation that often not all entities found in a text can be linked to a knowledge base. This creates a gap in coverage between what is in the text and what can be annotated with fine grained types.
This paper presents an approach to detect entity types using DBpedia type information and distributional semantics. The distributional semantics paradigm assumes that similar words occur in similar contexts. We exploit this by comparing entities with an unknown type to entities for which the type is known and assign the type of the most similar set of entities to the entity with the unknown type. We demonstrate our approach on seven different named entity linking datasets.
To the best of our knowledge, our approach is the first to combine word embeddings with external type information for this task. Our results show that this task is challenging but not impossible and performance improves when narrowing the search space by adding more context to the entities in the form of topic information.

1 Introduction

Fine grained entity typing facilitates precise queries to structured datasets. It can, for example, be used to query for all politicians or presidents in a dataset. With natural language processing techniques (NLP) becoming more accurate, structured datasets are increasingly being generated from text. However, there is still a gap between the results generated by most NLP techniques and what semantic web resources can offer.

Named entity recognition and classification (NERC) systems usually only discern a limited number of coarse grained types such as person, location, organisation and miscellaneous (CoNLL, [27]) or person, organisation, location, facility, weapon, vehicle and geo-political entity (ACE, [1]).

To obtain fine grained entity types for an entity, a named entity linking step is often employed to link recognised entities in an existing knowledge base such

as DBpedia. Thereby linked entities are enriched with the types of the resource, resolving the problem of not being able to perform fine grained queries. However, entity linking does not solve the entire problem, as not all entities can be linked to the knowledge base, for example, because there is no suitable resource present (often denoted as 'NIL' entities) or the resource may not contain any useful information about the entity to facilitate semantic querying [5].

In this paper, we focus on predicting the entity type of an entity regardless of its presence or absence in the knowledge base. Once the entity type has been established, a schema can be assigned to an entity which can serve as input for identifying other characteristics of the entity for example in a knowledge base population task. To tackle this task, we present an approach that employs distributional semantics and DBpedia types. We evaluate our approach on seven different entity linking benchmark datasets and make our resulting datasets available as a first dataset of NIL entities with fine grained type information. The contributions of this paper are threefold:

1. a method and implementation for fine grained entity typing;
2. quantitative and qualitative evaluation and analysis of the system on seven benchmark datasets; and
3. a new dataset for NIL entities including entity types.

The remainder of this paper is organised as follows. In Section 2, background and related work is discussed. Our approach is presented in Section 3. Section 4 describes the resources and datasets used for the experiments presented subsequently in Section 5. An analysis of the results (Section 6) and conclusions and future work (Section 7) wrap up this paper. All code, links to datasets and experiments are available via `https://github.com/MvanErp/entity-typing`.

2 Background and Related Work

Named entity recognition and classification has a long tradition in the natural language processing field, starting with the Message Understanding Conferences that were organised by DARPA between 1987 and 1997 [7]. The field also received a boost with the CoNLL 2002 and 2003 named entity recognition shared tasks [24,27], whose annotated datasets are still widely used for training and testing named entity recognition and classification approaches. However, the entity types used in these shared tasks and in the ACE challenges [1] are quite limited; CoNLL only distinguishes four entity types, and ACE seven main types as well as some subtypes. The main reason for this is that most systems developed for these tasks rely on supervised machine learning, for which sufficient examples of each entity type are needed. Some experiments with more elaborate type hierarchies [25] and (semi-)supervised machine learning for fine grained entity typing have been carried out [18], but this has not caught on much. Most likely due to the prohibitive expense of creating training datasets for this.

Named entity disambiguation or named entity linking systems can implicitly provide fine grained entity types. With Wikipedia and later DBpedia generic

and large resources were for the first time widely available. [13] present a system that uses the Wikilinks to link Wikipedia articles to relevant keywords in a text. [16] present the first entity linking system. Their system identifies entities in a text by a machine learning algorithm and then links them to a Wikipedia article, augmenting the text with the Wikipedia article's category information. For an overview of more entity linking approaches using Wikipedia see [8].

In the Semantic Web community entity linking systems such as [12,28] rely on the knowledge base providing a good coverage of the entities in a text to link. If the entity in the text does not have a suitable resource in the knowledge base, the system returns a NIL value at best, and at worst an incorrect link.

Approaches that deal with NILs or can be considered entity typing without entity linking are found in the domain of entity clustering [4]. However, these approaches generally do not leverage type information or hierarchies from external resources. Clustering methods and distributional models such as word2vec (which will be further explained in Section 4) have in common that they utilise the context surrounding a word or entity. Thus far, distributional models have been used for a wide variety of natural language processing tasks including relationship learning [19] and named entity recognition [26], but to the best of our knowledge, our work is the first to apply it to the entity typing.

In the remainder of the paper, we make a distinction between entity mentions, i.e. the textual reference in a text that denote entities where entities are things in the real world, oftentimes denoted by URIs in a knowledge base, e.g. a DBpedia resource. Entity types are properties of entities that express a categorisation of the entity. An example of two entity mentions referring to the same entity would be "Royal Air Force" and "RAF" both referring to `http://dbpedia.org/resource/Royal_Air_Force`.

3 Entity Typing using Distributional Semantics

Our approach for entity typing relies on the assumption that similar words[1] occur in similar contexts. As entities may not be mentioned often in a text, a large corpus of texts is also needed.

For the word embeddings in the model, we use word2vec [14][2], a popular word embeddings implementation. The idea behind word vectors is not new [2], but Mikolov et al. propose two new optimised architectures, in addition to a freely available implementation of the algorithm, making it possible to perform experiments with decently-sized datasets on fairly standard machines.[3]

The approach relies on a neural network to learn word vectors from a large text corpus, the idea being that similar words occur in similar contexts which can be captured by a word vector model. Such a model can be used to compute the semantic distance between two words or phrases, as well as algebraic operations on the vectors, an often mentioned example here being *vector("King")*

[1] As entities are made up of words, we hypothesise that this also extends to entities

[2] `https://code.google.com/archive/p/word2vec/`

[3] For this paper, we ran experiments on a Ubuntu machine with 2 CPUs, 16GB of RAM and most experiments did not take longer than 2 hours.

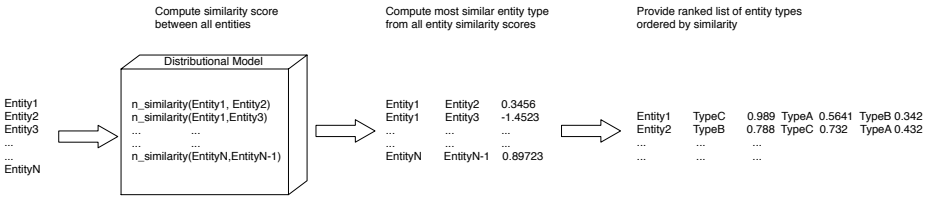

Fig. 1. Entity typing using distributional semantics system setup

- *vector("Man") + vector("Woman")* resulting in *vector("Queen")* being the closest response[15]. It also allows to compute a measure of semantic similarity between two words or groups of words which is what we will employ here. One expects the similarity between the entities 'George Bush' and 'Barack Obama' to be higher than between the entities 'George Bush' and 'Mexico City', which is indeed what the GoogleNews model returns (0.50 vs. 0.15). By performing this computation over many entity pairs and aggregating the entity types of the most similar entities, we aim to assign a meaningful entity type to an entity for which the entity type is unknown.

Figure 1 shows the system setup. A list of entity mentions for which the entity type is not known serves as input to the system. This entity mention is compared to all other entity mentions in the dataset for which the type information is available and the similarity between the two entities is computed. Thus, we average the scores of all entity types and produce a ranked list of the entity types whose entities are most similar to the entity at hand.

By using entity linking benchmark datasets that contain links to DBpedia, the entity types of the entities can be retrieved. Furthermore, the seven different datasets that will be described in the next section provide a wide range of entity mentions and types to evaluate various aspects of the approach.

Two types of experiments will be carried out: **1. Dataset-based experiments:** in this series of experiments, every entity mention within a dataset will be compared to all other entity mentions in that dataset. **2. Topic-based experiments:** in this series of experiments, the dataset is first split into topics and every entity mention is compared to only those entity mentions that are also in the same topic.

4 Resources and Datasets

The experiments carried out for this paper rely on existing algorithms and datasets which are described here.

4.1 Benchmark Entity Linking Datasets

We chose to test our approach on a number of freely available entity linking benchmark datasets, and were previously collected and described in [6]. Each of these datasets has different characteristics, that may present our approach with

Table 1. General statistics benchmark datasets

Dataset	Number of Entity Mentions	Number of Unique Entities	Number of NILs	Number of Types
AIDA-YAGO	11,862	5,029	4,333	195
2014 NEEL	3,084	2,081	0	205
2015 NEEL	5,346	2,643	1,699	213
OKE2015	773	501	116	55
RSS500	874	369	449	97
WES2015	9,753	6,016	1	211
Wikinews	1,724	251	660	48

different challenges, all have in common that are linked to DBpedia enabling us to leverage the type information from DBpedia.[4]

Table 1 displays some general statistics on the datasets. For the RSS500 dataset and WES2015, entities that did not have a DBpedia link were counted as NILs (these were linked to a dataset specific resource, but the type information for these was not available). In 2014 NEEL, NILs are not annotated.

AIDA-YAGO2 Dataset

The AIDA-YAGO2 dataset [9][5] is an extension of the most commonly used named entity recognition benchmark dataset, namely the CoNLL 2003 entity recognition task dataset [27]. The CoNLL 2003 dataset is based on a 10-day subset of Reuters news articles published between August 1996 and August 1997 by Reuters, to which part-of-speech and chunk tags were added automatically and named entities were added manually.

For this paper, we have mapped the Wikipedia URL to its corresponding DBpedia URI. Furthermore, the Reuters topic descriptions were reinserted into the articles in order to perform a series of experiments with topic classifications. The majority of the codes was added semi-automatically by Reuters; first a rule-based system proposes a topic, this is then checked by one or two human annotators. Next, the topic codes were cleaned up and its ancestors in the hierarchy were added through the process described in [11], whose corrected dataset we used.[6]

2014 and 2015 NEEL

The 2014 and 2015 Named Entity rEcognition and Linking (NEEL) dataset is made up of two Twitter datasets used in two consecutive challenges. The 2014 NEEL dataset [3][7] consists of 3,504 tweets extracted from over 18 million tweets provided by the Redites project. The tweets were collected over a period of 31 days between 15 July 2011 and 15 August 2011 and include noteworthy events.

[4] AIDA-YAGO2 originally contained Wikipedia URLs but these have been mapped to their corresponding DBpedia URIs

[5] `https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/aida/downloads/`

[6] Available from: `http://www.jmlr.org/papers/volume5/lewis04a/lyrl2004$_$rcv1v2$_$README.htm` Last visited: 27 April 2016

[7] `http://scc-research.lancaster.ac.uk/workshops/microposts2014/challenge/index.html`

The 2014 Microposts challenge dataset was created to benchmark automatic extraction and linking entities.

The 2015 NEEL corpus [21][8] is an extension of the 2014 dataset containing 6,025 tweets. Additional tweets published in 2013 were added to the original dataset. The resulting corpus was further extended to include entity types and NIL references. Entity references are linked to DBpedia resources.

OKE2015

The Open Knowledge Extraction Challenge 2015 (OKE2015) [20][9] corpus consists of 197 sentences from Wikipedia articles. Besides entities linked to DBpedia, the entities are also annotated with Dolce Ultra Lite classes,[10], coreference relations, and semi-automatic anaphora resolution, and detection of emerging entities. The corpus was split into a train and test set containing 96 sentences for the training set, and 101 for the test set.

RSS-500-NIF-NER

The RSS-500 dataset [23][11] contains data from 1,457 RSS feeds, including major international newspapers, covering a wide variety of topics. 500 sentences were chosen from an initial corpus of 11.7 million sentences and annotated by one researcher. The chosen sentences contain a formal relation (e.g. "..who was born in.." for dbo:birthPlace), that should occur more than 5 times in the 1% corpus.

WES2015

The WES2015 dataset was originally created to benchmark information retrieval systems [29].[12] The documents originate from a blog about history of science, technology, and art.[13] in which the entities are linked to DBpedia resources. The dataset also includes 35 annotated queries inspired by the blog's query logs, and relevance assessments between queries and documents. These were not used in the experiments described in this paper.

WikiNews/MEANTIME

The WikiNews/MEANTIME (hereafter referred to as 'Wikinews') [17].[14] is a linguistically and semantically annotated corpus of 120 news articles from the open news website Wikinews.[15] This corpus is divided into four sub-corpora: Airbus, Apple, General Motors and Stock Market. These are annotated with entities in text, including links to DBpedia, events, temporal expressions and

[8] http://scc-research.lancaster.ac.uk/workshops/microposts2015/challenge/index.html
[9] https://github.com/anuzzolese/oke-challenge
[10] http://stlab.istc.cnr.it/stlab/WikipediaOntology/
[11] https://github.com/AKSW/n3-collection
[12] http://yovisto.com/labs/wes2015/wes2015-dataset-nif.rdf
[13] http://blog.yovisto.com/
[14] http://www.newsreader-project.eu/results/data/wikinews
[15] https://en.wikinews.org/

semantic roles. This set of articles was selected to represent domain entities and events from the financial aspect of the automotive industry. The corpus is available in English, Spanish, Italian and Dutch. In our experiments, we limit ourselves to the English part of the corpus.

4.2 Word2vec models

In this paper, we use three different word2vec models, the first two are pre-trained models: 1) GoogleNews-vector-negative300.bin.gz[16] trained on part of the Google News dataset (∼100 billion words) [15][17] and 2) English Wikipedia (Feb 2015).[18] The third model was generated from the Reuters RCV1 corpus,[19] consisting of news wire published between August 1996 and August 1997. One of the main entity linking datasets described in the next section is derived from this dataset, therefore we chose to do an experiment involving this dataset and compare it to the Google News corpus. For all experiments, we use the Python gensim implementation of word2vec.[20]

5 Experiments and Results

The system outputs a ranked list of entity types for each entity mention. When measuring the performance of the system against the gold standard entity type, the precision at positions 1, 5 and 10 in the ranked results list is measured.

The results are divided into coarse- and fine-grained results. For the coarse-grained results, we only looked at the top level entity types in DBpedia, e.g. Agent, Place, Name, TopicalConcept etc. For the fine-grained results, we only looked at the most specific types in the DBpedia ontology. e.g. Airline, BeautyQueen, Monastery etc. This is quite a strict evaluation metric as we only either the most generic or most specific exact entity type per entity. If a system returns entity types from elsewhere in the type hierarchy, these are not currently not considered in the aggregated results, but these will be discussed in the qualitative analysis in Subsection 6.3.

Tables 2 and 3 present the results of the experiments of the dataset-based experiments. The first table displays the percentage of correct entity types returned by the system in the ranked list at positions 1, 5 and 10. Table 3 provides statistics on the coverage of the entity mentions in the various word2vec models.

For the AIDA-YAGO and Wikinews datasets, a topic classification of the articles from which the entities are derived is available too. In this subsection, a series of experiments is described in which the entity datasets are further divided into datasets by topic. As this results in fewer entity comparisons (as only entities within a topic are compared), this narrows down the search space.

[16] `https://drive.google.com/file/d/0B7XkCwpI5KDYNlNUTTlSS21pQmM/edit?usp=sharing`

[17] Unfortunately, no further information about the Google News corpus is available as it is not an open dataset

[18] `https://github.com/idio/wiki2vec`

[19] `http://trec.nist.gov/data/reuters/reuters.html`

[20] `https://radimrehurek.com/gensim/models/word2vec.html`

Table 2. Results per entity dataset as percentage of correct types returned by the system in position 1, 5 or 10.

Dataset	GoogleNews						Wikipedia						Reuters					
	Coarse			Fine			Coarse			Fine			Coarse			Fine		
	@1	@5	@10	@1	@5	@10	@1	@5	@10	@1	@5	@10	@1	@5	@10	@1	@5	@10
AIDA-YAGO2	0.00	0.23	0.42	0.26	5.71	16.62	0.00	0.19	0.38	0.38	4.33	14.06	0.00	0.12	0.25	0.40	5.86	16.24
2014 NEEL	0.00	1.53	3.45	0.13	5.62	12.48	0.00	1.25	3.26	0.04	4.99	12.42	0.00	0.76	2.93	0.05	2.42	9.18
2015 NEEL	0.00	1.77	3.05	0.04	5.60	12.79	0.00	1.57	2.69	0.08	4.64	12.40	0.00	0.53	1.32	0.05	1.95	6.37
OKE2015	0.00	0.84	1.05	1.68	7.58	16.84	0.00	0.76	1.33	2.66	7.02	17.27	0.00	0.00	1.55	4.88	9.98	20.40
RSS500	0.00	0.12	0.25	6.37	9.36	15.23	0.00	0.00	0.27	4.28	7.35	12.83	0.00	0.00	0.33	0.17	2.97	7.26
WES2015	0.00	0.61	1.91	0.10	2.19	5.41	0.00	0.97	3.60	0.01	5.51	9.48	0.00	1.02	3.25	0.05	3.57	6.74
Wikinews	0.00	8.39	16.46	1.90	11.87	28.48	0.00	16.52	22.71	2.32	8.77	26.71	0.00	6.00	12.69	1.91	8.05	25.24

Table 3. Statistics on coverage of entities in the different models

	Total #	GoogleNews		Wikipedia		Reuters	
	Entity mentions	Found	Not Found	Found	Not Found	Found	Not Found
AIDA-YAGO2	11,862	9,103	2,759 (23.26%)	8,294	3,568 (30.07%)	8,937	2,925 (24.66%)
2014 NEEL	3,084	2,347	737 (24.18%)	2,487	597 (19.36%)	1,982	1,102 (35.73%)
2015 NEEL	5,346	2,949	2,397 (44.83%)	2,885	2,461(46.03%)	2,101	3,245 (60.70%)
OKE2015	773	554	219 (28.33%)	620	153(19.79%)	529	244 (31.57%)
RSS500	874	801	73 (8.35%)	748	126(14.41%)	606	268 (30.66%)
WES2015	9,753	6,743	3,010 (30.86%)	8,278	1,475 (15.12%)	6,496	3,257 (33.40%)
Wikinews	1,724	985	739 (42.86%)	1,311	413 (23.96%)	1,265	459 (26.62%)

As described in Section 4, we inserted the Reuters topics classification into the AIDA-YAGO dataset. Two series of experiments were run: one with only the top level Reuters topics (AIDA-YAGO Coarse) and one with the more fine grained Reuters topics (AIDA-YAGO Fine). In total, the Reuters topics classification set contains top level 23 topics, but only 21 are present in the dataset. In total, the Reuters hierarchy contains 103 subtopics, of which 68 are present in the dataset. One article does not have a topic ascribed to it, this article was treated as a separate topic resulting in 22 topics in total in the coarse grained top-level Reuters topics and 69 topics in the finer grained topics experiments. An overview of the topics and their distribution over the AIDA-YAGO dataset can be found on our github page.

The entities in the Wikinews topics are fairly evenly spread, the largest topic contains 275 unique entities and the smallest 104, with a median at 140. The AIDA-YAGO topics are quite diverse in nature and in division; in the fine grained topic division, the smallest topics only contains 7 unique entity mentions (Personal Income) and (Reserves), and the largest 18,616 (Sports). The median lies around 108 entity mentions per topic. Even on the more coarse grained topic division the differences are large: the smallest topic contains 29 entity mentions (Management), whereas the largest contains 30,320 (Government/Social), but the median lies around 295 entity mentions per topic.

The Wikinews dataset is divided into four themes, by treating the entities in each of these themes, we can also experiment with a really high-level topic classification. Figure 2 and Table 4 show the results of the experiment series in which entity linking was performed within a topic. For space considerations,

Table 4. Results per entity dataset with aggregated topics

Dataset	GoogleNews						Wikipedia						Reuters					
	Coarse			Fine			Coarse			Fine			Coarse			Fine		
	@1	@5	@10	@1	@5	@10	@1	@5	@10	@1	@5	@10	@1	@5	@10	@1	@5	@10
AIDA-YAGO Coarse	0.00	0.27	6.42	0.02	7.80	22.28	0.00	0.30	5.29	0.04	6.73	20.24	0.00	0.16	2.83	0.02	7.04	21.60
AIDA-YAGO Fine	0.00	0.31	10.37	0.30	11.64	29.61	0.00	0.45	10.08	0.18	10.01	27.06	0.00	0.24	6.92	0.27	9.57	25.40
Wikinews Topics	0.00	4.23	14.37	0.28	3.10	9.01	0.00	3.35	11.51	0.21	2.09	6.49	0.00	0.00	6.36	3.64	7.50	10.23

only the results on the AIDA-YAGO dataset are presented here, the full results can be found on the `https://github.com/MvanErp/entity-typing`.

In Figure 2, the graphs on the lefthand side show the results on the coarse grained Reuters topics, and the righthand side show the results on the fine grained Reuters topics. The figures shown here only present the results on the most specific DBpedia types (columns 'Fine' in Table 4).

6 Results Analysis and Discussion

In this section, we first present a quantitative analysis of the dataset- and topic-based experiments, followed by a qualitative analysis of a data sample.

6.1 Dataset-based experiments

At first sight, the results may indicate that the approach does not yield the desired results for entity typing but there are two main reasons that indicate otherwise. Firstly, the smaller datasets OKE2015 and Wikinews perform better than the larger datasets. When considering the entities in these datasets, they do not only cover fewer different entities, but also fewer different entity types (see Table 1). The topic-based experiments demonstrate that datasets that are less broad, i.e. centred around a particular topic are better suited to our approach. Secondly, to preserve coherence, we only focused on type information from the DBpedia ontology (`http://dbpedia.org/ontology/`) but not all resources contain a DBpedia ontology type, which will be discussed in Subsection 6.3.

Let us first consider the differences in results between the different word2vec models. The GoogleNews model is almost 10x larger than the Reuters model (3.4GB vs 374MB) and for the majority of the datasets tested, it yields a better performance, due to there being more contexts in the model for a particular entity. However, as Table 2 shows, this does not hold for the AIDA-YAGO dataset, whose entities are extracted from the Reuters corpus. The results show that there is a clear advantage to using a model based on the data the entities are derived from, as this ensures that the original context in which the entities were mentioned are also encoded in the model.

Another interesting observation is that the approach consistently ranks more fine grained entity types first over more coarse grained types. This accounts for the coarse grained Score@1 columns yielding scores of 0.00, whilst the approach manages to find the most fine grained entity type (according to the gold standard) in the top position in some cases. This is a positive signal as the goal is to provide fine grained entity types. It should be noted here that intermediate

entity types are not taken into account in this analysis, e.g. in the type hierarchy Place - PopulatedPlace - Settlement - Village, the coarse grained evaluation measured whether the type 'Place' was returned by the system, and the fine grained analysis whether the type 'Village' was returned by the system. As the hierarchy does not have a fixed number of levels, measuring the performance at each step is difficult to aggregate and is left for future work.

As Table 3 indicates, the corpus used to find and compare entities is an important factor in the experiments. Not surprisingly, the entities in RSS500 are largely covered by the GoogleNews model, with only 8.35% of the entity mentions missing, but this figure jumps to 30.66% for the Reuters model. Both corpora cover news, but the Reuters corpus dates from nearly 20 years ago, when different entities played a role in the news. This is also apparent from the coverage on the 2014 and 2015 NEEL datasets in Reuters, and to a lesser extent in GoogleNews and Wikipedia. Tweets generally have a different style than news with more abbreviations, capitals and hash tags and Twitter handles. As the models and entity mentions are not normalised, this yields fewer matches.

As the OKE2015 dataset is based on Wikipedia, the coverage by the Wikipedia model on this dataset is higher than that of the news models. The coverage of the WES2015 dataset is also better in the Wikipedia model than the news models, this is due to the WES2015 dataset covering science topics. Mentions of pre-socratic philosophers such as 'Anaxagoras' and 17th century botanists such as 'Nicholas Culpeper' are simply less frequent in the news than in sources such as Wikipedia. In the word2vec models, there is also a slight bias towards more frequent entities, it would simply not be possible to capture all hapaxes as this makes the model less efficient. In training the Reuters model, only words that occur at least 10 times in the corpus were taken into account, as the same parameters were used as those for the GoogleNews model. Further experiments with different parameter settings may also positively influence the performance.

6.2 Topics-based experiments

When looking at the fine grained results, on the righthand side of Figure 2, a few topics yield scores of 0. Topics 1 (Advertising Promotion), 58 (Reserves), and 62 (Share listings) prove difficult for the Wikipedia model to classify correctly. These topics contain 18, 7 and 14 entity mentions respectively. However, Topic 40 (Leading indicators), with only 20 entity mentions, does yield reasonable scores. Upon inspecting the entities in this topic, this is probably due to these being fairly generic entities such as 'Hungary', 'Budapest' and 'Spain', these entities occur so often that only a few contexts suffice to type them. Interestingly, the GoogleNews model also has trouble with Topic 3 (Art Culture Entertainment), which contains 256 entity mentions. As the entities are derived from the Reuters corpus, it performs decently there, but also the Wikipedia model does well here as it contains a fair amount of information regarding entertainment [10]. Topics 21 (EC External Relations) and 22 (EC Monetary Economic) with 57 and 12 entities perform well across all models. Topics 29 (Fashion) and 48 (Money Supply) obtain a 100% score on the top 10 results in the Wikipedia and GoogleNews models respectively, providing a hunch about the coverage of the model.

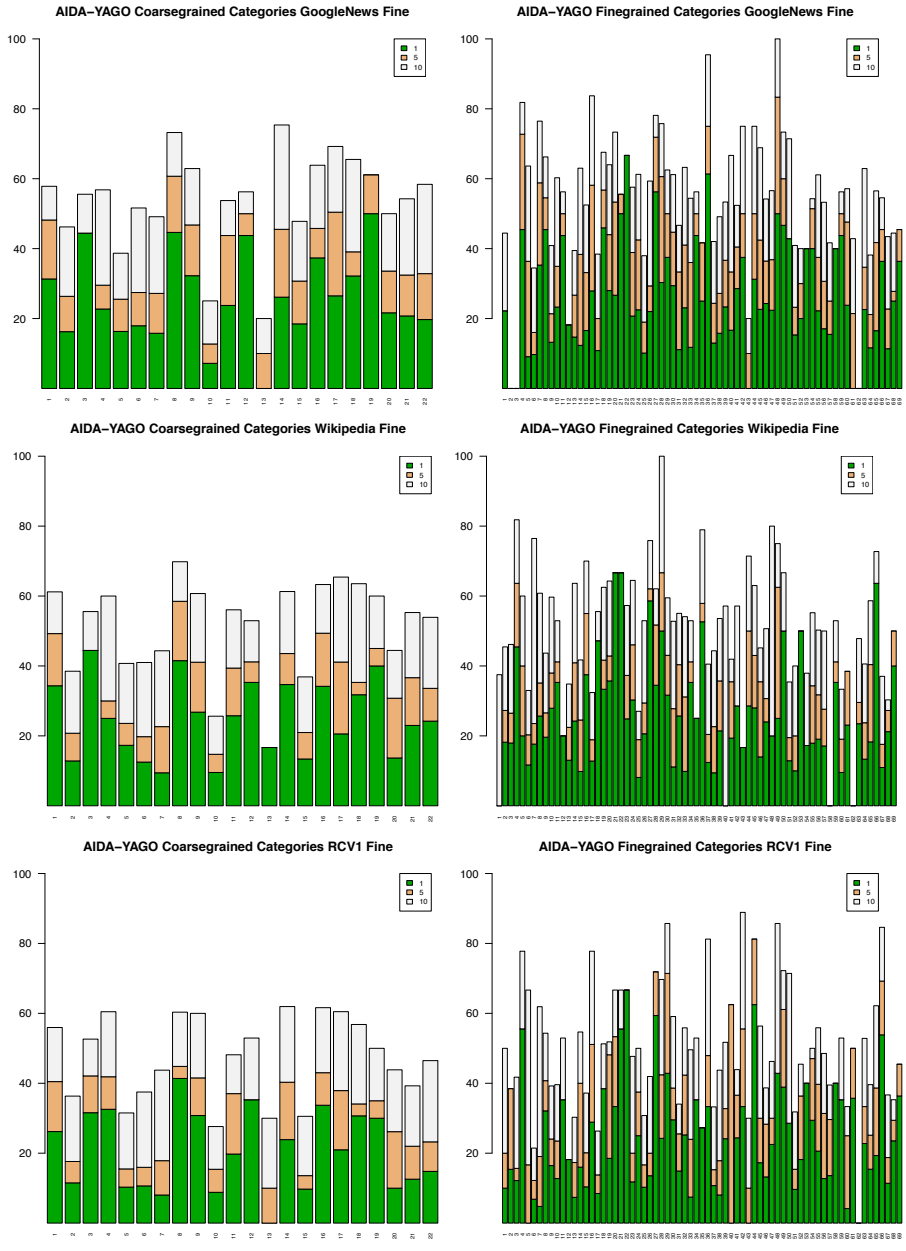

Fig. 2. Results on AIDA-YAGO dataset with entities divided into topics

For the coarse grained topics, on the lefthand side of Figure 2, the size of the topic plays a role again. Topic 13 (Management) is the smallest topic, with only 29 entity mentions, and it yields the lowest results in all models. However, bigger is not always better, as Topic 10 (Government/Social) with 30,320 entity mentions confirms.

This topic contains sports as well as subtopics such as obituaries, elections, weather, crime and fashion. The topics that stand out positively here are 3 (Consumer finance) and 19 (Output capacity) with 37 and 29 entity mentions, respectively, and 4 (Contracts/Orders), 8 (European Community) and 14 (Markets/Marketing) with 105, 279 and 407 entity mentions respectively.

The figures for the more coarse grained DBpedia types can be found on the `Github page`. These are largely in line with the fine grained results where the same topics present the different models with challenges. As with the overall dataset results, the system aims for more specific entity types first, and less specific types are typically found in the rank 5-10 range and hardly any coarse grained entity types are found at rank 1.

The results for the `Wikinews dataset` do not improve much in the per-topic setting. This dataset is not that large, and still quite broad. Another issue in this dataset is that there are many entities that are not in DBpedia, thus resulting in fewer entities with context typing, this is partly due to the wide variety in entities accepted in the annotations such as 'GM's package' and 'JSF contract'. The fine grained DBpedia type results on the Wikinews dataset are a bit 'all or nothing' in the sense that the majority of the correct results are found at rank 1, proportionally fewer correct results are found lower in the ranked list. As with the overall dataset results, the system aims for more specific entity types first, and less specific types are typically found in the rank 5-10 range.

6.3 Qualitative Analysis

For the qualitative analysis we looked at various samples of the data to get a more in-depth understanding of the mismatch between the system output and the gold standard types.

Not matching the Gold Standard We took a random sample of 200 entries dataset-based experiments for which no match was found between the gold standard types and the entity types proposed by the system. There are four interesting observations here:

1. Multifaceted entities: Some entities are quite difficult to capture in a single DBpedia type, e.g. `http://dbpedia.org/resource/Thomas_Erle` is classified as a `MilitaryPerson` in the gold standard. He was indeed an army general, but later a politician who sat in the House of Commons and our system returns type suggestions such as `Politician`, `MemberOfParliament` and `Congressman`. Other ambiguous entities are found when the entity mention can have several meanings. E.g. "Buffalo" is linked to `http://dbpedia.org/resource/Buffalo,_New_York` which is of course a `City`, but the system returns types such as `Eukaryote`, `Species` and `Animal`. To resolve these, more context may be included in the query.

2. Type Specificity: As the system favours more specific entity types over more generic ones, it may in some cases suggest a type that is also correct, but was not present in the gold standard. For example `http://dbpedia.org/resource/Justin_Bieber` has type `Person` in DBpedia, and the system suggests `MusicalArtist`.

3. Ambiguous entity mentions: Some entity mentions are very difficult to classify without additional context. The entity "massacre" is linked to `http://dbpedia.org/resource/The_Massacre` which is a music album in DBpedia. However, the system predicts types such as `MilitaryConflict` and `Event`. Other difficult entities to type for the system are thing such as 'month' and numbers, which are annotated in some of the datasets. For such 'generic' terms denoting entities, the distributional semantics approach is too coarse and in some cases it is debatable whether the item should have been annotated as an entity at all.

4. Gold standard limitations: No dataset is perfect and due to their size it is impossible even for the very active Wikipedia and DBpedia communities to check every resource but there are some types in the gold standard that are at least a bit puzzling such as `http://dbpedia.org/resource/KFC` which is a `SportsTeam` according to the gold standard and the `http://dbpedia.org/resource/US_Open_(tennis)` which is a `PopulatedPlace`.

No gold standard DBpedia types available Some entity mentions have a DBpedia link assigned to them, indicating that they are present in the knowledge base, but the gold standard does not contain any DBpedia types to evaluate the system output against. There are two main reasons for this.

1. Choice of ontology: For reasons of coherence and manageability of the evaluations, only DBpedia ontology types were considered in our experiments. But DBpedia resources may also have Yago, Umbel, Geonames, Schema, and Wikidata types assigned to them. Although the majority of the resources in our datasets have at least one DBpedia type assigned, there are also resources that for example only have Yago types (e.g. `http://dbpedia.org/resource/BRIC`).

2. Redirects: The entity mention 'China' is linked to `http://dbpedia.org/resource/People's_Republic_of_China` which only contains yago types. However, this resource actually also has a redirect link to `http://dbpedia.org/page/China` which does contain the DBpedia types `Place` and `Country` for which were also suggested by the system.

Redirects and other ontologies were not considered in these experiments, but follow-up research could include these.

NILs To gain insights in the performance of the approach on NIL entities, we manually evaluated the performance of the Reuters model on 200 random entity mentions from the AIDA-YAGO dataset with fine grained Reuters classes. As this is not a formal annotation, we only considered whether a 'reasonable' entity type was suggested by the system at rank 1, or within ranks 5 or 10, but we did not specify whether this was a fine grained or coarse grained type.

In 31 of the cases, our method suggests a reasonable entity type at position 1 in the ranked list. In an additional 60 cases, a relevant entity type was found within the first 5 results, and in another 20 cases within the first 10 positions.

As there seem to be many Cricketers in the dataset, these tend to be classified correctly. The system also returns some very specific correct entity types in certain cases such as `BusCompany` for Swebus AB.

As the system favours more specific entities, it often comes up with various different types of sports teams (`RugbyTeam`, `SoccerTeam`) or athlete types (`TennisPlayer`, `RugbyPlayer`), so the suggested entity types are in the correct domain. Some entity mentions are very difficult to type, such as "Ontario-based", "US-led" and "non-EU". There are also some mentions of divisions of companies such as "Sydney Newsroom". Whilst strictly speaking not a company in itself, it is part of a company and thus deemed reasonably typed by the system.

There are also some cases where the entity mention boundary was probably not correct in the dataset such as "British Airways-American", although the system does return `Airline` as a type suggestion.

The analysis shows that the system suggests very reasonable entity types for the NILs in these datasets. An inspection of a number of topics and suggested types shows that the topic boundaries effectively limit the number of entity types for a topic, which is very helpful to the system as it provides a stronger context to type entities from. The typed NILs datasets generated in these experiments are available through our website.

The analyses described in this section indicate that some of the system suggestions are more reasonable than the results in Table 2 suggest.

7 Conclusion and Future Work

In this paper, we presented a novel method and experiments for fine grained entity typing using distributional semantics and DBpedia. Our results show that this is a difficult task but when entity mentions are limited within a topic, the system achieves reasonable performance. We tested this topic-based approach on two datasets with available topic information. In future experiments, we aim to use topic detection to investigate different topic granularity levels on all datasets to gain insights into the optimal topic-entity type-entity mention ratio.

Moreover, we evaluated our approach using three different word embedding models on seven different benchmark datasets. Our quantitative and qualitative analyses show that the performance of the system depends on the size and domain of the dataset. Ideally, these language models are trained on in-domain texts. Luckily, this is quite feasible as the models do not require annotated data. Normalising the datasets and entity mention queries may also boost coverage.

Not all issues can be resolved by carefully tuning the experiment parameters. The fact that often the approach does suggest a relevant entity type within the first 10 types also presents interesting avenues of research for post-processing.

One of the goals of our research is to discover more information about entities that are not present in a given knowledge base or for which the knowledge base does not contain sufficient relevant information to reason with. An entity typing step can provide us with likely entity types, a subsequent relation extraction may be used to further rerank the entity types. For non-NIL entities, we can also investigate whether having a fine grained entity type available may help improve entity linking.

Overall, our method provides a promising first step in using implicit and explicit domain knowledge for entity typing.

Acknowledgements

The research for this paper was made possible by the CLARIAH-CORE project financed by NWO: http://www.clariah.nl.

References

1. ACE (Automatic Content Extraction) english annotation guidelines for entities. http://www.ldc.upenn.edu/Projects/ACE/ (2006)
2. Bengio, Y., Ducharme, R., Vincent, P., Jauvin, C.: A neural probabilistic language model. Journal of Machine Learning Research 3, 1137–1155 (2003)
3. Cano, A.E., Rizzo, G., Varga, A., Rowe, M., Stankovic Milan, Dadzie, A.S.: Making Sense of Microposts (#Microposts2014) Named Entity Extraction & Linking Challenge. In: 4^{th} International Workshop on Making Sense of Microposts. #Microposts (2014)
4. Elsner, M., Charniak, E., Johnson, M.: Structured generative models for unsupervised named-entity clustering. In: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL'09). pp. 164–172 (2009)
5. van Erp, M., Ilievski, F., Rospocher, M., Vossen, P.: Missing mr. brown and buying an abraham lincoln - dark entities and dbpedia. In: Proceedings of NLP & DBpedia 2015 workshop in conjunction with 14th International Semantic Web Conference (ISWC2015), CEUR Workshop Proceedings (2015)
6. van Erp, M., Mendes, P., Paulheim, H., Ilievski, F., Plu, J., Rizzo, G., Waitelonis, J.: Evaluating entity linking: An analysis of current benchmark datasets and a roadmap for doing a better job. In: Proceedings of LREC 2016. (2016), preprint available from: https://mariekevanerp.files.wordpress.com/2012/06/evaluating-entity-linking-1.pdf
7. Grishman, R., Sundheim, B.M.: Message understanding conference - 6: A brief history. In: Proceedings International Conference on Computational Linguistics (1996)
8. Hachey, B., Radford, W., Nothman, J., Honnibal, M., Curran, J.R.: Evaluating entity linking with wikipedia. Artificial Intelligence 9, 130–150 (2013)
9. Hoffart, J., Yosef, M.A., Bordin, I., Fürstenau, H., Pinkal, M., Spaniol, M., Taneva, B., Thater, S., Weikum, G.: Robust Disambiguation of Named Entities. In: Conference on Empirical Methods in Natural Language Processing. EMNLP (2011)
10. Kittur, A., Chi, E.H., Suh, B.: What's in wikipedia? : mapping topics and conflict using socially annotated category structure. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'09). pp. Pages 1509–1512. ACM, New York, NY, USA (2009)
11. Lewis, D.D., Yang, Y., Rose, T.G., Li, F.: RCV1: A new benchmark collection for text categorization research. Journal of Machine Learning Research 5, 361–397 (2004)
12. Mendes, P.N., Jakob, M., García-Silva, A., Bizer, C.: Dbpedia spotlight: Shedding light on the web of documents. In: Proceedings of the 7th International Conference on Semantic Systems (I-SEMANTICS 2011). ACM New York, NY, USA, Graz, Austria (Sept 7-9 2011)

13. Mihalcea, R., Csomai, A.: Wikify! linking document to encyclopedic knowledge. In: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management (CIKM'07). pp. 233–242 (2007)
14. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. In: arXiv preprint arXiv:1301.3781 (2013)
15. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Proceedings of NIPS (2013)
16. Milne, D., Witten, I.H.: Learning to link with wikipedia. In: Proceedings of the 17th ACM conference on Information and knowledge management (CIKM'08). pp. 509–518 (2008)
17. Minard, A.L., Speranza, M., Urizar, R., na Altuna, B., van Erp, M., Schoen, A., van Son, C.: MEANTIME, the newsreader multilingual event and time corpus. In: Proceedings of the 10th edition of the Language Resources and Evaluation Conference (LREC 2016) (2016)
18. Nadeau, D.: Semi-supervised named entity recognition: learning to recognize 100 entity types with little supervision. Ph.D. thesis, University of Ottawa (2007)
19. Nguyen, T.H., Grishman, R.: Relation extraction: Perspective from convolutional neural networks. In: Proceedings of NAACL-HLT 2015. pp. 39–48. Denver, Colorado, USA (May 31 - June 5 2015)
20. Nuzzolese, A.G., Gentile, A.L., Presutti, V., Gangemi, A., Garigliotti, D., RobertoNavigli: Open knowledge extraction challenge. In: Semantic Web Evaluation Challenges (2015)
21. Rizzo, G., Cano Amparo E, Pereira, B., Varga, A.: Making sense of Microposts (#Microposts2015) named entity recognition & linking challenge. In: 5^{th} International Workshop on Making Sense of Microposts. #Microposts (2015)
22. Rizzo, G., Troncy, R.: NERD: A Framework for Unifying Named Entity Recognition and Disambiguation Extraction Tools. In: 13^{th} Conference of the European Chapter of the Association for computational Linguistics (EACL'12) (2012)
23. Röder, M., Usbeck, R., Hellmann, S., Gerber, D., Both, A.: N3-a collection of datasets for named entity recognition and disambiguation in the nlp interchange format. In: 9^{th} Language Resources and Evaluation Conference. LREC (2014)
24. Sang, E.F.T.K.: Introduction to the conll-2002 shared task: Language-independent named entity recognition. In: Proceedings of CoNLL-2002. Taipei, Taiwan (2002)
25. Sekine, S., Sudo, K., Nobata, C.: Extended named entity hierarchy. In: Proceedings of the Third International Conference on Language Resources and Evaluation. pp. 1818–1824 (2002)
26. Sienčnik, S.K.: Adapting word2vec to named entity recognition. In: Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015). pp. 239–243. Vilnius, Lithuania (May 11-13 2015)
27. Tjong Kim Sang, E.F., De Meulder, F.: Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In: Conference on Computational Natural Language Learning. CoNLL (2003)
28. Usbeck, R., Ngomo, A.C.N., Röder, M., Gerber, D., Coelho, S.A., Auer, S., Both, A.: AGDISTIS -graph-based disambiguation of named entities using linked data. In: Proceedings of the 13th International Semantic Web Conference (ISWC 2014). pp. 457–471. Riva del Garda, Italy (October 2014)
29. Waitelonis, J., Exeler, C., Sack, H.: Linked Data Enabled Generalized Vector Space Model to Improve Document Retrieval. In: Proceedings of NLP & DBpedia 2015 workshop in conjunction with 14th International Semantic Web Conference (ISWC2015), CEUR Workshop Proceedings (2015)