"One Entity per Discourse" and "One Entity per Collocation" Improve Named-Entity Disambiguation

Ander Barrena*, Eneko Agirre*, Bernardo Cabaleiro, Anselmo Peñas**, Aitor Soroa***

*IXA NLP Group / University of the Basque Country, Basque Country
abarrena014@ikasle.ehu.es, e.agirre@ehu.es, a.soroa@ehu.es

**UNED NLP & IR Group, Madrid
anselmo@lsi.uned.es, bcabaleiro@lsi.uned.es

Abstract

The "one sense per discourse" (OSPD) and "one sense per collocation" (OSPC) hypotheses have been very influential in Word Sense Disambiguation. The goal of this paper is twofold: (i) to explore whether these hypotheses hold for entities, that is, whether several mentions in the same discourse (or the same collocation) tend to refer to the same entity or not, and (ii) test their impact in Named-Entity Disambiguation (NED). Our experiments show consistent results on different collections and three state-of-the-art NED system. OSPD hypothesis holds in around 96%-98% of documents whereas OSPC hypothesis holds in 91%-98% of collocations. Furthermore, a simple NED post-processing in which the majority entity is promoted, produces a gain in performance in all cases, reaching up to 8 absolute points of improvement in F-measure. These results show that NED systems would benefit of considering these hypotheses into their implementation.

1 Introduction

The "one sense per discourse" (OSPD) hypothesis was introduced by Gale et al. (1992), and stated that a word tends to preserve its meaning when occurring multiple times in a discourse. They estimated that the probability of two occurrences of the same polysemous noun drawn from one document having the same sense to be around 94% for documents from Grolier encyclopedia, and 96% for documents from Brown, based on word senses from the Oxford Advanced Learner's Dictionary and a handful of examples. A few years later, Krovetz (1998) reported 66% on larger corpora (SemCor and DSO) annotated with WordNet senses by third parties, but, unfortunately, he only reported how many polysemous nouns occurred with a single sense in **all** documents, not in each document. In the context of statistical machine translation, Carpuat (2009) reported that, 80% of the time, words occurring multiple times in a source document are translated into a single word in the target language.

In the case of entities, OSPD is closely related to coreference, where the task is to find whether two different mentions (perhaps using different surface strings like "John" and "he") in a document refer to the same entity or not. For instance, the coreference system presented by (Lee et al., 2013), uses a heuristic which links mentions in a document that share the same surface string: "This sieve [heuristic] accounts for approximately 16 CoNLL F1 points improvement, which proves that a signicant percentage of mentions in text are indeed repetitions of previously seen concepts". Our paper actually quantifies the amount of those repetitions for entities, providing additional evidence for the heuristic.

The "one sense per collocation" (OSPC) hypothesis was introduced by Yarowsky (1993), stating that a word tends to preserve its meaning when occurring with the same collocate. Yarowsky tested his hypothesis for several definitions of collocate, including positional collocates (word to left or right) and syntactic collocations (governing verb of object, governing verb of subject, modifying adjective). He reported entropy on train data, as well as disambiguation performance on unseen data, with the precision ranging between 90% and 99% for a handful of words with two distinct homograph senses, like, e.g. "bass" or "colon". In larger-scale research, Martinez and Agirre (2000) measured the precision

Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers,
pages 2260–2269, Dublin, Ireland, August 23-29 2014.

> **Abbott** Beefs Up Litigation Reserves NORTH CHICAGO, Ill. (AP) **Abbott** Laboratories Inc., bracing for a costly settlement in a federal investigation involving the prostate-cancer drug Lupron, said Friday it was increasing litigation reserves by $344 million. As part of the announcement, **Abbott** said it had restated its quarterly results and is now reporting a loss of $319.9 million for the first three months of this year rather than a profit. The move comes amid long-running negotiations between the U.S. Department of Justice and TAP Pharmaceutical Products, the 50-50 joint venture between **Abbott** and Takeda Chemical Industries of Japan that made Lupron. **Abbott** said in January ...

Figure 1: Example of OSPD for entities. All occurrences of "**Abbott**" refer to "Abbott Laboratories".

of similar collocations on corpora (Semcor and DSO) annotated by third parties with finer-grained senses from WordNet, reporting lower figures around 70%.

In this paper, we take a collocation to be a word (or multiword term) that co-occurs with the target named-entity more often than would be expected by chance. In our case we use syntactic dependencies to extract co-occurring terms.

These two hypotheses have been very influential, and have inspired multiple heuristics and methods in Word Sense Disambiguation research (Agirre and Edmonds, 2007, Chapters 5,7,10,11). In this work we are going to show that both hypotheses hold for named-entities as well, and that the hypotheses can be used to post-process the output of any Named-Entity Disambiguation system (NED) to improve its performance. NED, also known as Entity Linking, takes as input a named-entity mention in context and assigns it a specific entity from a given entity repository (Hachey et al., 2012; Daiber et al., 2013).

In the first part of this work we are going to test whether the two hypotheses hold for entity mentions with respect to a repository of entities extracted from Wikipedia. For instance, do all occurrences of mention "Abbott" in a document refer to the same entity? Do all occurrences of mention "CPI" as subject of verb "rise" refer to the same entity? Do all occurrences of "CDU" in relation to "Merkel" refer to the same entity? The examples in Figures 1 and 2 show evidence that this is indeed the case. The experiments aim at quantifying in which degree OSPD and OSPC hypotheses hold for entities[1].

In the second part of the paper, we will explore a simple method to incorporate OSPD and OSPC hypotheses to any existing NED system, showing their potential. After running the NED system, we take its output and observe, for each mention string, which is the entity returned most often for a given document (or collocation), assigning to all occurrences the majority entity. We tested the improvements with a freely available NED system (Daiber et al., 2013), a reimplementation of a strong Bayesian NED system (Han and Sun, 2011) and an in-house graph-based system. We got statistically significant improvements for all systems and "one sense" hypotheses that we tested, with a couple exceptions.

In order to check the OSPD and OSPC hypotheses for entities, we first looked into existing datasets. AIDA (Hoffart et al., 2011)[2] is a publicly available hand-tagged corpus based on the CoNLL named-entity recognition and disambiguation task dataset. AIDA contains links of all entity mentions in full documents, so it is a natural fit for OSPD. We estimated OSPD based on more than 4,000 mentions that occur multiple times in a document. For completeness, we also estimated OSPD at the collection level.

OSPD and OSPC are independent of each other, as one is applied at the document level and the other at the corpus level, focusing on the entities that occur with a specific collocation. Multiple occurrences of a target string in a document usually occur with different collocations, and conversely, multiple occurrences of a target string with a specific collocation typically occur in different documents. Note also that singletons (entities that are only mentioned once in a document) are not affected by OSPD, but could be affected by OSPC.

In order to estimate OSPC, no available corpus existed, so we decided to base our dataset on the TAC KBP 2009 Entity Linking dataset[3] (TAC2009 for short) (Ji et al., 2010). The TAC2009 dataset involves 138 mention strings, which have been annotated in several documents drawn primarily from Gigaword[4].

[1] For the sake of clarity we will also refer to OSPD and OSPC for entities as OSPD and OSPC.

[2] http://www.mpi-inf.mpg.de/yago-naga/aida/downloads.html

[3] http://www.nist.gov/tac/2013/KBP/EntityLinking/index.html

[4] http://catalog.ldc.upenn.edu/LDC2003T05

CPI subject-of rise:

China's consumer price index, or **CPI**, rose 2.8 percent last December.
In the 10 months to October, the **CPI** rose 1.35 percent, the core price index grew 1.13 percent ...
Measured on a month-on-month basis, March **CPI** rose 2.3 percent from February, ...
... still lower than in China, Hong Kong and Singapore, whose **CPI**s have rised 8.0 percent, ...
The core **CPI** rose 0.2 percent, in line with Wall Street expectations.

Angela Merkel has **CDU**:

... who share power with Merkel's **CDU** nationally in an uneasy " grand coalition " ...
Economy Minister Michael Glos, also from the CSU, the sister party to Merkel's **CDU** ...
In the past Merkel's **CDU** had been able to rely on the CSU's strength in Bavaria ...
... but while her conservative **CDU** wanted new legal tools to do so, ...
The new development has put a further strain on Merkel's **CDU** ...

Figure 2: Examples of OSPC for entities, showing five examples for a syntactic collocation (top row) and fie examples for a more specific proposition (bottom row). "**CPI**" might refer to "Comunist Party of India" or "Consumer Price Index", among others, but refers to the second in all cases. "**CDU**" can refer to the German "Christian Democratic Union" or "Catholic Distance University", among others, but refers to the first in all cases.

We extracted several syntactic collocations for those 138 mention strings from Gigaword, and hand-annotated them, yielding an estimate for the OSPC. Note that TAC2009 only provides the annotation for a specific mention in a document, so we had to annotate by hand the rest of occurrences in the documents. For instance, we analyzed examples of "CPI" as subject of the verb "rise" (cf. Figure 2). Some of the syntactic collocations like the subjects of verb "has" seemed very uninformative, so we decided to also check the OSPC hypothesis on more specific collocations, involving more complete argument structures. For instance, we checked "ABC" occurring as subject of "has" with object "radio". We call this more specific collocations *propositions* (Peñas and Hovy, 2010).

The paper is structured as follows. We will first present the resources used in this study. Section 3 presents the results of OSPD. Section 3.1 extends OSPD when, instead of documents, we take the complete collection. Section 4 presents the study of OSPC both for syntactic dependencies and propositions. Section 5 presents the experiments where OSPD and OSPC are used to improve the performance of existing systems. Finally, we draw the conclusions and future work.

2 Resources used

AIDA is based on the corpus used in the CONLL named-entity recognition and classification task, where all entities in full documents had been linked to the referred Wikipedia articles (using the 2010 Wikipedia dump). We use the full AIDA dataset, with 1,393 documents, 34,140 disambiguated entity mentions, where 27,240 are linked to a Wikipedia article. All in all there are 6,877 distinct mention strings (types) which are linked at least once to a Wikipedia article. The rest refer to articles not in Wikipedia (NIL instances), and were discarded. This corpus covers news from a sample of a few days spanning from 1996-05-28 to 1996-12-07.

In order to prepare our dataset for OSPC, we chose the dataset of the TAC KBP 2009 Entity Linking competition, as this dataset have been extensively used in Entity Linking evaluation. In addition, the corpus used in the task was very large, allowing us to mine relevant collocations (see below). We manually annotated the occurrences in the extracted collocations, producing two datasets, one for each kind of collocation (cf. Section 4). Note that the TAC KBP organizers only annotated one specific mention in each target document. For completeness, we also tagged the rest of the occurrences of the target mentions in the documents, thus allowing us to provide OSPD estimated based on TAC2009 data as well. This is the third dataset that we annotated by hand. The hand-annotation was performed by a single person, and later reviewed by the rest of the authors. The three annotation datasets are publicly available[5]. Hand-

[5] http://ixa2.si.ehu.es/OEPDC

NHasN	”U.S. dollar”
NPN	”condition of anonymity”
NVN	”official tells AFP”
NVNPN	”article maintains interest within layout”
NVPN	”others steal from input”
VNPN	”includes link to website”

Table 1: List of the six patterns used to extract propositions, with some examples.

tagging is costly, so we tagged around 250 examples of syntactic collocations and around 250 examples of propositions.

Note that both AIDA and TAC2009 contain mentions that were not linked to a Wikipedia article because the mention referred to an entity which was not listed in the entity inventory. We ignored all those cases (called NIL cases), as we would need to investigate, for each NIL, which actual entity they refer to.

The collocations were extracted from the TAC KBP collection (Ji et al., 2010), comprising 1.7 million documents, 1.3 millions from newswire and 0.5 millions from the web. We have parsed them with the Stanford CoreNLP software (Klein and Manning, 2003), obtaining around 650 million dependencies (De Marneffe and Manning, 2008). We selected subject, object, prepositional complements and adjectival modifiers as the source for syntactic collocations. In order to provide more specific collocations, we implemented the syntactic patterns proposed in (Peñas and Hovy, 2010), which produce so-called propositions. The result is a database with 16 million distinct propositions. Table 1 shows the six patterns used in this work, together with some examples.

In order to know whether a mention is ambiguous, we built a dictionary based on Wikipedia which lists, for each string mention, which entities it can refer to. We followed the construction method of (Spitkovsky and Chang, 2012), which checked article titles, redirects, disambiguation pages and hyperlinks to find mention strings that can be used to refer to entities. Contrary to them, we could not access hyperlinks in the web, so we could use only those in Wikipedia. According to our dictionary, the ambiguity of the mentions that we are studying is very high, 26.4 entities on average for the mentions in AIDA, and 62.6 entities on average for the mentions in TAC2009.

3 One entity per discourse

In order to estimate OSPD we divided the number of times a mention string referred to different entities in the document with the number of times a mention string occurred multiple times in the document. In the denominator and numerator we count each mention-document pair once.

Regarding AIDA, we found 12,084 occurrences of mentions which occurred more than once in a document, making 4,265 unique mention-document pairs[6] (cf. Table 2). In the vast majority of the cases those mentions refer to a single entity in the document, and only in 170 cases the mentions in the document refer to several entities. The last row in Table 2 shows the ratio between those values, 96.01%, showing that OSPD is strong in this dataset.

We also checked OSPD in the TAC2009 dataset. Out of the 138 distinct mention strings used in the task, we discarded those only linked to NIL (that is, no corresponding Wikipedia article existed) and those which were not ambiguous (that is, they had only one entity in the dictionary, cf. Section 2). That leaves 105 mention strings, occurring 1,776 times in 918 different documents, which we annotated by hand. The 105 strings occurred 1,776 times in 918 documents. Removing the cases where the mention occurred only once, we were left with 1,173 occurrences, which make 334 unique mention-document pairs, of which only 6 occurred with more than one sense (rightmost row in Table 2). This yields an estimate for OSPD of 98.2%.

[6] By unique mention-document pairs we mean that we only count once for a mention occurring multiple times in a document. For instance if mention *Smith* occurs 10 times in the whole corpus, 8 times in document *A* and 2 times in document *B*, we count two unique mention-document pairs.

	AIDA	TAC2009
Mention-document pairs	4,265	334
Ambiguous pairs	170	6
OSPD	**96.0%**	**98.2%**

Table 2: One entity per discourse: per document statistics in AIDA and TAC2009 datasets. Pairs stand for the number of unique mention-document pairs. The 4,265 pairs in AIDA correspond to 12,084 occurrences of mentions, and the 334 pairs in TAC2009 correspond to 1,173 occurrences.

	All mentions		First mention	
	AIDA	TAC2009	AIDA	TAC2009
Mention types	3,363	105	2,731	105
Ambiguous types	475	26	454	25
OSPD (collections)	**85.9%**	**75.2%**	**83.4%**	**76.2%**

Table 3: One entity per collection: statistics in AIDA and TAC2009. In the first two columns ("All mentions") we consider all mention types ($3,363$ types in AIDA correspond to $23,726$ occurrences of mentions, and 105 types in TAC2009 correspond to $1,776$ occurrences). In the second two columns ("First mention") we leave only the first mention of each document (in this case, there are $2,731$ mention types in AIDA which correspond to $15,275$ occurrences, and 105 types in TAC2009 corresponding to 941 occurrences).

Finally, we also thought about measuring OSPD on the Wikipedia articles, where many mentions have been manually linked to their respective article. Unfortunately, we noted that Wikipedia guidelines explicitly prevent authors linking a mention multiple times: *Generally, a link should appear only once in an article, but if helpful for readers, links may be repeated in infoboxes, tables, image captions, footnotes, and at the first occurrence after the lead*[7]. The fact that Wikipedia editors did not explicitly state exceptions to the above rule (e.g. for cases where the word or phrase is used to refer to two different articles, thus breaking the OSPD hypothesis) is remarkable, and might indicate that Wikipedia editors had not felt the need to challenge the OSPD hypothesis.

3.1 One entity per collection

We took the opportunity to also explore "one entity per collection", which gives an idea of what is the spread of entities for whole document collections. In this case, there is no need to count mention-document pairs, as there is one single document, the collection, so we estimate the hypothesis according to mention types. The first two columns in table 3 shows that, overall, mentions which occurred more than once in the collection tend to refer to the same entity 85.9% of the time in AIDA, and 75.2% of the time in TAC2009.

As we know that multiple mentions in a document tend to refer to one entity, the second two columns in table 3 offers the statistics when factoring out multiple occurrences of mention in a document, that is, leaving the first mention in each document. The statistics are very similar, with minor variations.

We think that the lower estimate for TAC2009 is an artifact of how the TAC KBP organizers set up the dataset, as they were explicitly looking for cases where the target string would refer to different entities, making the task more challenging for NED systems. This fact does not affect OSPD for documents, as those strings still tend to refer to a single entity per document, but given the need to find occurrences for different entities, the organizers (Ji et al., 2010) did focus on strings occurring with different entities across the document collection. This is in contrast with AIDA, where they tagged all named-entities occurring in the target documents. Had the organizers of TAC2009 focused on a random choice of strings and documents, the one entity per collection would also hold to the high degree exhibited in AIDA, as the genre of most of the documents is also news (as in AIDA).

[7] http://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style/Linking#What_generally_should_be_linked

	Syn. coll.	Propositions
Mention-collocation pairs	58	61
Ambiguous pairs	5	1
OSPC	**91.4%**	**98.4%**

Table 4: One entity per collocation: statistics for syntactic collocations and propositions. The 58 mention-collocation pairs correspond to 262 occurrences, and the 61 mention-proposition pairs to 279.

4 One entity per collocation

In order to estimate OSPC for **syntactic collocations**, we manually annotated several occurrences of the 138 mention strings of the TAC2009 dataset. Hand-tagging mention entities is a costly process, so we chose (at random) one syntactic dependency relation for each of the 138 mention strings that occurred more than five times in the corpus. We then hand-tagged at random five occurrences of each collocation (cf. Figure 2). This method would provide a maximum of 5 examples for each of the 138 mentions, but after checking the minimum frequency of the collocations, the quality of the context, repeated sentences, mentions that are not ambiguous in the dictionary, and whether the mention could be attached to an entity in the database, the actual number was lower. All in all we found 58 mention-collocation pairs (262 occurrences) for syntactic collocations (cf. middle column in Table 4). Only 5 mentions referred to more than one entity per collocation, yielding that OSPC for syntactic collocation is around 91.4%.

To gather the dataset for **propositions**, we followed the same method as for the syntactic collocations, that is, we chose (at random) one propositions involving one of the 138 mention strings that occurred more than five times in the corpus, and hand-tagged at random five occurrences of each proposition (cf. Figure 2. As with syntactic collocations, we also found a limited number of mentions filling the desired properties. That left 61 mention-collocation pairs (279 occurrences) for propositions (cf. right column in Table 4). Only 1 mention referred to more than one entity per proposition, yielding OSPC for propositions around 98.4%. This shows that the more specific the context is, the stronger is the link between mention and entity.

5 Improving performance

In order to check whether any of the "one sense" hypothesis above could improve the performance of a NED system, we followed a simple procedure: After running the NED system, we take its output and observe, for each mention string, which is the entity returned most often for a given document (or collocation), assigning to all occurrences the majority entity. In case of ties, we return the entity with the highest support from the NED system. We tested the improvements on three NED systems: the freely available DBpedia Spotlight, a reimplementation of a strong Bayesian NED system and a graph-based system.

DBpedia Spotlight is a freely available NED system (Daiber et al., 2013), based on a generative probabilistic model (Han and Sun, 2011). Nowadays it is one of the most widely used NED systems and attains performances close to state-of-the-art (Daiber et al., 2013)We used the default values of the parameters for all the experiments in this paper.

We also tested an in-house reimplementation of the generative probabilistic model presented in (Han and Sun, 2011). This is a state-of-the-art system which got the same accuracy as the best participant (72.0) when evaluated in the non-NIL subset of TAC2013.

UKB is a freely-available system for performing Word Sense Disambiguation and Similarity based on random walks on graphs (Agirre and Soroa, 2009). Instead of using it on WordNet, we represented Wikipedia as a graph, where vertices are the wikipedia articles and edges represents bidirectional hyperlinks among Wikipedia pages, effectively implementing a NED system. We used a Wikipedia dump from 2013 in our experiments. UKB is a competitive, state-of-the-art system which attained a score of 69.0 when evaluated in the non-NIL subset of the TAC2013 dataset.

The input of the systems is the context of each mention to be disambiguated, in the form of a 100 token window centered in the target mention. In NED, the identification of the correct mention to be

Mention in context	Entity
Abbott Beefs Up Litigation ...	Abbot_Kinney
Abbott Laboratories Inc., bracing ...	Abbott_Laboratories
Abbott said it had restated ...	Abbott_Laboratories
venture between **Abbott** and Takeda ...	Abbott_Laboratories
Abbott said in January ...	Abbott_Laboratories

Figure 3: Applying OSPD: Each of the five occurrences of Abbott in the document in Figure 1 has been tagged independently by a NED systems, which return the correct entity in all but one case (precision 80%). Applying OSPD would return the correct entity (Abbott_Laboratories) in all cases, improving precision to 100%.

	AIDA			TAC 2009		
	Prec.	**Recall**	**F1**	**Prec.**	**Recall**	**F1**
Spotlight	83.24	63.90	72.30	64.48	46.44	53.99
+ OSPD Discourse	**84.17**	70.01	76.44	**64.65**	**48.50**	**55.42**
+ OSPD Collection	84.02	**74.64**	**79.05**	56.24	47.98	51.78
UKB	70.09	69.03	69.55	67.70	67.64	67.67
+ OSPD Discourse	71.30	70.23	70.76	**70.21**	**70.21**	**70.21**
+ OSPD Collection	**75.79**	**74.64**	**75.21**	68.84	68.84	68.84
(Han and Sun, 2011)	65.71	65.11	65.41	65.49	65.49	65.49
+ OSPD Discourse	67.77	67.37	67.57	66.27	66.27	66.27
+ OSPD Collection	**74.29**	**73.89**	**74.09**	**68.24**	**68.24**	**68.24**

Table 5: Applying OSPD: NED performance on AIDA and TAC2009 OSPD datasets, including each of the three NED systems, and the results after applying OSPD at the document and collections levels. Bold marks best result for each system.

disambiguated is part of the problem. AIDA does provide gold mentions, but TAC2009 only provides a query string which might be just a substring of the real mention in the document. We treated both corpus in the same way. In the case of DBpedia Spotlight we use the built-in mention spotter. In the case of our in-house implementations, we use the longest string that matches a valid entity mention in the system, as given by the dictionary (cf. Section 3).

Some of the NED systems do not return an entity for all mentions, so we evaluate precision, recall and the harmonic mean (F1 measure). Statistical significance has been estimated using Wilcoxon. We reused the same corpora as in the previous sections for the evaluation, and also removed all NIL mentions (i.e. mentions which refer to an entity not in Wikipedia).

5.1 One entity per discourse

We report the improvements using OSPD for both **document** and **collection** levels. At the document level, we relabel mentions that occur multiple times in a document using the entity returned most times by the NED system in that document. Figure 3 illustrates the idea for a NED system on the same sample document as in Figure 1. At the collection level, we relabel mentions using the entity returned most times by the NED systems in the whole collection.

Table 5 reports the results of the performance as evaluated on mentions occurring multiple times in the AIDA and TAC2009 datasets. The numbers in the left part of the table correspond to the performance as evaluated on mentions occurring multiple times in AIDA documents. Note that the number of occurrences where OSPD at the collection level can be applied is larger (a superset of those for OSPD at the document level), as, for instance, a mention string occurring once in three different documents won't be affected by OSPD at the document level, but it could be relabeled at the collection level. We were especially interested in making the numbers between OSPD at the document and collection levels

CPI subject-of rise	Angela Merkel has CDU:
Consumer_price_index	Christian_Democratic_Union_(Germany)
Consumer_price_index	Catholic_Distance_University
Communist_Party_of_India	Christian_Democratic_Union_(Germany)
Communist_Party_of_India	Christian_Democratic_Union_(Germany)
Consumer_price_index	Christian_Democratic_Union_(Germany)

Figure 4: Applying OSPC: A NED system system tagged each example in Figure 2 independently. For CPI, the precision is 60%, but after relabeling with OSPC it would be 100%. For CDU, the improvement is from 80% to 100%.

	Syntactic collocations			Propositions		
	prec.	**recall**	**F1**	**prec.**	**recall**	**F1**
Spotlight	82.46	66.41	73.57	74.67	60.22	66.67
+ OSPC	**82.63**	**67.18**	**74.11**	**74.79**	**62.72**	**68.23**
UKB	75.86	75.57	75.72	67.87	67.38	67.63
+ OSPC	**78.54**	**78.24**	**78.39**	**68.59**	**68.10**	**68.35**
(Han and Sun, 2011)	75.57	75.57	75.57	71.33	71.33	71.33
+ OSPC	**78.24**	**78.24**	**78.24**	**73.12**	**73.12**	**73.12**

Table 6: Applying OSPC: NED performance on TAC2009, including each of the three NED systems, and the results after applying OSPC for syntactic collocations and propositions. Bold is used for best results for each system.

directly comparable, and therefore report the results on the same occurrences, that is, the occurrences where OSPD at the document level can be applied.

The results show a small but consistent improvement for OSPD at the document level in precision, recall and F1 for the three NED systems, around 1 or 2 absolute points. The improvements when applying OSPD at the collection level are also consistent, but remarkably larger, between 5 and 9 absolute points. All improvements are statistically significant (p-value below 0.01).

Table 5 also reports the results after applying OSPD to TAC2009 instances which occurred more than once in a document. Results for OSPD at document level and collection level follow the same methodology as for AIDA. The improvement at the collection level is not so consistent, with a loss in performance for Spotlight, a small improvement for UKB, and a larger improvement for (Han and Sun, 2011). All differences across the table are statistically significant (p-value below 0.01).

While the OSPD at the document level is strong in both corpora, Section 3.1 showed that the OSPD at the collection level is only strong in AIDA, with a much lower estimate in TAC2009. This fact would explain why the improvement with OSPD at the collection level is not consistent. Following the rationale in Section 3.1, we think that had the organizers of the task chosen strings and documents at random, the improvement in TAC 2009 at the collection level would be also as high as in AIDA. The high improvement in AIDA at the collection level compared to the more modest improvement at the document level, despite having a lower OSPD estimate (cf. Section 3.1), could be caused by the fact that there are more occurrences and evidence in favor of the majority entity.

5.2 One entity per collocation

Figure 4 shows the application of OSPC to the output of a NED system to two sample collocations in our dataset. In this case, the application of OSPC would increase precision to 100%. The actual result on the datasets produced in Section 4 for syntactic collocations and propositions is reported on table 6.

Regarding syntactic collocations, table 6 shows that the improvement is small but consistent for the three systems on precision, recall and F1, ranging from 0.5 to 2.5 absolute points in F1 score. The results for propositions also show the same trend, with consistent improvements across the table. All differences in the two tables are statistically significant (p-value < 0.01), except for UKB.

6 Conclusions and future work

Our study shows that OSPD holds for 96%-98% (in the AIDA and TAC2009 datasets, respectively) of the mentions that occur multiple times in documents. We also measured OSPD at the collection level (86% and 75%, respectively). OSPC holds for 91% of the mentions that occur multiple times in the syntactic collocations that we studied, and 98% of the mentions that occur multiple times in more specific collocations. We reused the publicly available AIDA dataset for estimating OSPD. In addition, we created a dataset to study OSPC based on the TAC KBP Entity Linking 2009 task dataset, which is publicly available[8].

We carefully chose to estimate both OPSD and OSPC on TAC2009, in order to make the numbers between OSPD and OSPC comparable. The OSPD numbers for AIDA are very similar to those obtained on TAC2009, providing complementary evidence. Although the high estimate of OSPD for entities was somehow expected, the high estimate of OSPC for the syntactic collocations, especially the propositions, was somehow unexpected, given the high ambiguity rate of the discussed strings, and the fact that the ambiguity included similar entities, like for instance ”ABC” which can refer, among other 190 entities, to the American Broadcasting Company or the Australian Broadcasting Corporation.

Our results also show that a simple application of the OSPD and OSPC hypotheses to the output of three different NED systems improves the results in all cases. Remarkably, the highest performance gain, 8 absolute points, was for OSPD at the collection level in the AIDA corpus.

The results presented here could be largely dependent on the domain and genre of the documents, as well as the definition of collocation. Our work is a strong basis for claiming that OSPD and OSPC hold for entities, but the evidence could be further extended exploring alternative operationalization of collocations and a larger breadth of genres and domains.

For the future we would like to check whether these hypotheses can be further used to improve current NED systems. The OSPD hypothesis can be used to jointly disambiguate all occurrences of a mention in a document. The OSPC hypothesis could be used to acquire important disambiguation features, or to perform large-scale joint entity linking. The OSPD for whole collections could be useful for documents on specific domains, and for domain adaptation scenarios.

Acknowledgements

This work was partially funded by MINECO (CHIST-ERA READERS project – PCIN-2013-002- C02-01) and the European Commission (QTLEAP – FP7-ICT-2013.4.1-610516, OPENER – FP7-ICT-2011-SME-DCL-296451). Ander Barrena is supported by a PhD grant from the University of the Basque Country.

References

Eneko Agirre and Philip Edmonds. 2007. *Word Sense Disambiguation: Algorithms and Applications*. Springer Publishing Company, Incorporated, 1st edition.

Eneko Agirre and Aitor Soroa. 2009. Personalizing pagerank for word sense disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, pages 33–41.

Marine Carpuat. 2009. One translation per discourse. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions*, DEW '09, pages 19–27, Stroudsburg, PA, USA. Association for Computational Linguistics.

Joachim Daiber, Max Jakob, Chris Hokamp, and Pablo N. Mendes. 2013. Improving efficiency and accuracy in multilingual entity extraction. In *Proceedings of the 9th International Conference on Semantic Systems*, I-SEMANTICS '13, pages 121–124, New York, NY, USA. ACM.

[8] `http://ixa2.si.ehu.es/OEPDC`

Marie-Catherine De Marneffe and Christopher D. Manning. 2008. Stanford typed dependencies manual. *URL http://nlp. stanford. edu/software/dependencies manual. pdf.*

William A. Gale, Kenneth W. Church, and David Yarowsky. 1992. One sense per discourse. In *Proceedings of the workshop on Speech and Natural Language*, HLT '91, page 233237, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ben Hachey, Will Radford, Joel Nothman, Matthew Honnibal, and James R. Curran. 2012. Evaluating Entity Linking with Wikipedia. *Artif. Intell.*, 194:130–150, January.

Xianpei Han and Le Sun. 2011. A Generative Entity-mention Model for Linking Entities with Knowledge Base. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pages 945–954.

Johannes Hoffart, Mohamed A. Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stephan Thater, and Gerdhard Weikum. 2011. Robust Disambiguation of Named Entities in Text. In *Conference on Empirical Methods in Natural Language Processing, Edinburgh, Scotland, United Kingdom 2011*, pages 782–792.

Heng Ji, Ralph Grishman, Hoa Trang Dang, Kira Griffitt, and Joe Ellis. 2010. Overview of the tac 2010 knowledge base population track. In *Third Text Analysis Conference (TAC 2010).*

Dan Klein and Christopher D Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 423–430. Association for Computational Linguistics.

Robert Krovetz. 1998. More than one sense per discourse. In *NEC Princeton NJ Labs., Research Memorandum.*

Heeyoung Lee, Angel Chang, Yves Peirsman, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2013. Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Computational Linguistics*, 39(4).

David Martinez and Eneko Agirre. 2000. One sense per collocation and genre/topic variations. In *Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics - Volume 13*, EMNLP '00, page 207215, Stroudsburg, PA, USA. Association for Computational Linguistics.

Anselmo Peñas and Eduard Hovy. 2010. Filling knowledge gaps in text for machine reading. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 979–987. Association for Computational Linguistics.

Valentin I. Spitkovsky and Angel X. Chang. 2012. A Cross-lingual Dictionary for English Wikipedia Concepts. *Eighth International Conference on Language Resources and Evaluation (LREC 2012).*

David Yarowsky. 1993. One sense per collocation. In *Proceedings of the workshop on Human Language Technology*, HLT '93, page 266271, Stroudsburg, PA, USA. Association for Computational Linguistics.