Giving a Sense: A Pilot Study in Concept Annotation from Multiple Resources

Roman Sudarikov and Ondřej Bojar

Charles University in Prague
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Malostranské náměstí 25, 11800 Praha 1, Czech Republic
http://ufal.mff.cuni.cz/
{sudarikov,bojar}@ufal.mff.cuni.cz

Abstract: We present a pilot study in web-based annotation of words with senses coming from several knowledge bases and sense inventories. The study is the first step in a planned larger annotation of "grounding" and should allow us to select a subset of these "dictionaries" that seem to cover any given text reasonably well and show an acceptable level of inter-annotator agreement.

Keywords: word-sense disambiguation, entity linking, linked data

1 Introduction

Annotated resources are very important for training, tuning or evaluating many NLP tasks. Equipped with experience in treebanking, we now move to resources for word sense disambiguation (WSD) and entity linking (EL). By EL, we mean the task of attaching a unique ID from some database to occurrences of (named) entities in text [1]. Both entity linking and word-sense disambiguation have been extensively studied, see for example [2–4]. Although only a few researches consider several knowledge bases and sense inventories at once [1, 5], the convergence between these two task is apparent, for example, the 2015 SemEval Task 13 promoted research in the direction of joint word sense and named entity disambiguation [6].

We understand the terms *ontology*, *knowledge base* and *sense inventory* in the following way:

- *Ontology* is a formal representation of a domain of knowledge. It is an abstract entity: it defines the vocabulary for a domain and the relations between concepts, but an ontology says nothing about how that knowledge is stored (as physical file, in a database, or in some other form), or indeed how the knowledge can be accessed.
- *Knowledge base* is a database, a repository of information that can be accessed and manipulated in some predefined fashion. Knowledge is stored in knowledge base according to an ontology.
- *Sense inventory* is a database, often build based on a corpus, and providing clustered *senses* for the words or expressions in the corpus.

However, we recognize the blending of *knowledge bases* and *sense inventories*, so we will use very generic terms *dictionary* or *resource* interchangeably for either of them.

In this pilot study, we examine several such *dictionaries* in terms of their coverage and annotator agreement. Unlike other works on "grounding", which try to link only the most important words in the sentence [7, 8], we aim at *complete* coverage of a given text, i.e. all content words or multi-word expressions regardless their part of speech or role in the sentence. Some of the examined resources have a clear bias towards some parts of speech, for example, valency dictionaries cover only verbs. We nevertheless ask our annotators to annotate even across parts of speech if the matching POS is not included in the resource. For instance, verbs can get nominal entries in Wikipedia and nouns get verb frames.[1]

In Section 2, we describe the sense inventories included in our experiment. Section 3 provides a unifying view on these sources and introduces our annotation interface. We conducted two experiments with English and Czech texts using the interface, slightly adapting interface for the second run. Details are in Section 4 and Section 5.

2 Resources Included

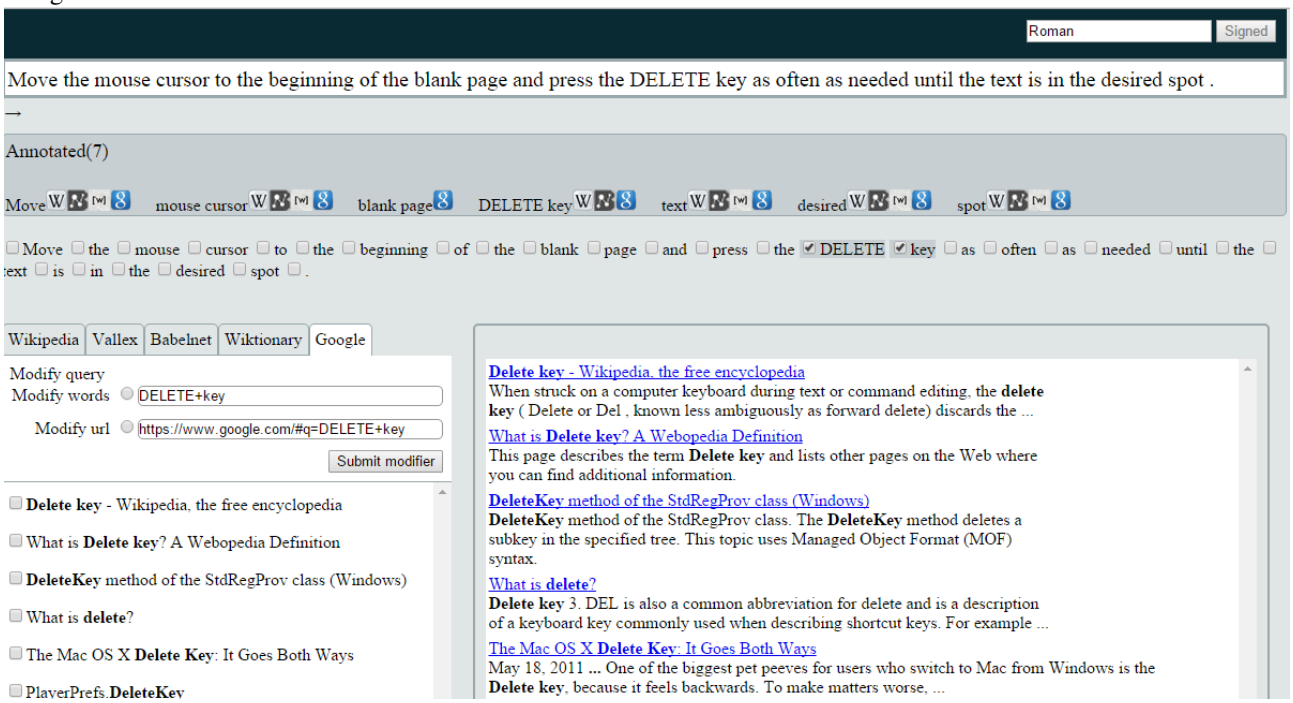
Sense inventories and knowledge bases are plentiful and they differ in many aspects including the domain coverage, level of detail, frequency of update, integration of other resources and ways of accessing them. Some of them implement Resource Description Framework, the metadata data model designed by W3C for the better data representation in Semantic Web, while others are simply collections of links in the web.

We selected the following subset of general *resources* for our experiment:

BabelNet [10] is a multilingual knowledge base, which combines several knowledge resources including Wikipedia, Wordnet, OmegaWiki and Wiktionary. The sources are automatically merged and accessible via offline Java API or online REST API. An added benefit is the multilinguality of BabelNet: the same resource can be used for genuine (as opposed to cross-lingual) annotation for both languages of our interest, English and Czech.

[1] The conversion of nouns to predicates whenever possible is explicitly demanded in some frameworks, e.g. in Abstract Meaning Representation (AMR, [9]).

Figure 1: Annotation interface, annotating the words “DELETE key” in the sentence “Move the mouse cursor. . . ” with Google Search “senses”.



The main limitation is that BabelNet is not updated continuously, so we also added both live Wikipedia and Wiktionary as separate sources. BabelNet provides information about nouns, verbs, adjectives and adverbs, but as stated above, we are interested also in cross-POS annotation.

Wikipedia[2] is currently the biggest online encyclopedia with live updates from (hundreds of) thousands of contributors so it can cover new concepts very quickly. Wikipedia tries to nest all possible concepts as nouns. For example, `en.wikipedia.org/wiki/funny` redirects to the page “Humour”.

Wiktionary[3] is a companion to Wikipedia that covers all parts of speech. It includes multilingual thesaurus, phrase books, language statistics. Each word in Wiktionary can have etymology, pronunciation, sample quotations, synonyms, antonyms and translations, for better understanding of the word.

PDT-VALLEX and EngVallex (Valency lexicons for Czech and English): Valency or subcategorization lexicons formally capture verb valency frames, i.e. their syntactic neighborhood in the sentence [11, 12]. We use the valency lexicons for Czech and English in their offline XML form as distributed with the tree editor TrEd 2.0[4].

Google Search[5] (GS): From our preliminary experiments, we had the impression that no resource covers all expressions seen in our data, but searching the web provides some explanation almost always. We thus include the top ten results returned by Google Search as a special kind of *dictionary*, where the “concept” is a query string and each result is considered to be its’ “sense”.

Aside from coverage and frequency of updates, another reason to include GS is that it provides “senses” at a very different level of granularity than others. For instance, the whole Wiktionary page can appear as one of the options in GS “senses”. It will also often be a very sensible choice, despite it actually covers several different meanings of the word.

We find the task of matching senses coming from different ontologies and providing a different angle of view or granularity very interesting. The current experiments serve as a basis for its further investigation.

3 Annotation Interface

To provide a unified view on the various resources, we use the terms *query*, *selection list* and *selection*. Given an expression in a text, which can be a word or a phrase, even a non-continuous one, and a resource which should be used to annotate it, the system construct a query. Querying the resource, we get a selection list, i.e. a list of possible senses.

The process of extracting the selection list depends on the resource. It is straightforward for Google Search (each result becomes an option) and complicated for Wiktionary,

[2] `http://wikipedia.org`
[3] `http://wiktionary.org`
[4] `http://ufal.mff.cuni.cz/tred/`
[5] `http://google.com`

Table 1: Selection statistics, the first (upper part) and second (lower part) annotation experiments

					One or more senses selected			
Source	Total	Whole page	Bad List	None	1	2	3	4 or more
Babelnet	28	-	1	3	23	1	0	0
Google Search	71	-	1	9	36	15	5	5
CS Vallex	38	-	0	2	29	6	1	0
EN Vallex	19	-	1	0	18	0	0	0
CS Wikipedia	38	-	9	12	15	1	0	1
EN Wikipedia	114	-	26	16	63	3	0	6
CS Wiktionary	21	-	1	3	7	4	5	1
EN Wiktionary	21	-	0	0	18	2	1	0
Babelnet	98	24	0	10	54	6	2	2
Google Search	93	0	0	26	19	16	11	21
EN Vallex	15	4	0	3	6	2	0	0
EN Wikipedia	103	23	7	36	35	2	0	0
EN Wiktionary	98	17	23	4	40	9	2	3

see Section 3.1 below. In principle and to include any conceivable resource, even field-specific or ad hoc ones, the annotator should be free to *select the selection list* prior to the annotation.

Our annotation interface allows to overwrite the query for cases where the automatic construction does not lead to a satisfactory selection list.

Finally, the annotator is presented with the selection list to make his choice (or multiple choices). Overall, the annotator picks one of these options:

Whole Page means that the current URL is already a good description of the sense and no selection list is available on the page. The annotators were asked to change the query and rather obtain a selection list (e.g. a disambiguation page in Wikipedia) whenever possible.

Bad List means that the extraction of selection list failed to provide correct senses. The annotators were supposed to try changing the query to obtain a usable list and resort to the "Bad List" option only if inevitable.

None indicates that the selection list is correct but that it lacks the relevant sense.

One or more senses selected is the desired annotation: The list, for the particular pair of *selected word(s)* and *selected resource*, was correct and the annotator was able to find the relevant sense(s) in the list.

Our annotation interface (Figure 1) shows the input sentence, tabs for individual sense inventories, the selection list from the current resource and also the complete page where the selection list comes from. The procedure is straightforward: (1) select one or more words in the sentence using checkboxes, (2) select a resource (we asked our annotators to use them all, one by one), (3) check if the selection list is OK and modify the query if needed, (4) make the annotation choice by marking one or more of the checkboxes in the selection list, and (5) save the annotation.

3.1 Queries and Selection Lists for Individual Resources

This is how we construct queries and extract selection lists for each of our *dictionaries* given one or more words from the annotated sentence:

BabelNet We search BabelNet for the lemma of the selected word (or the phrase of lemmas if more words are selected). The selection list is the list of all obtained BabelNet IDs.

Google Search We search for the lemmas of the selected words and return the snippets of the top ten results. The selection list is the list of snippets' titles.

Wikipedia We search for the disambiguation page for the selected words and, if not found, we search for the page with the title matching the lemmas of the selected words. The selection list for disambiguation pages is constructed by fetching hyperlinks appearing within listings nested in particular HTML blocks. For other pages we fetch links from the Table of Contents and the first hyperlink from each listing item.

Wiktionary We search for the page with the title equal to the lemmas of the selected words. The selection list is created using the same heuristics as for Wikipedia.

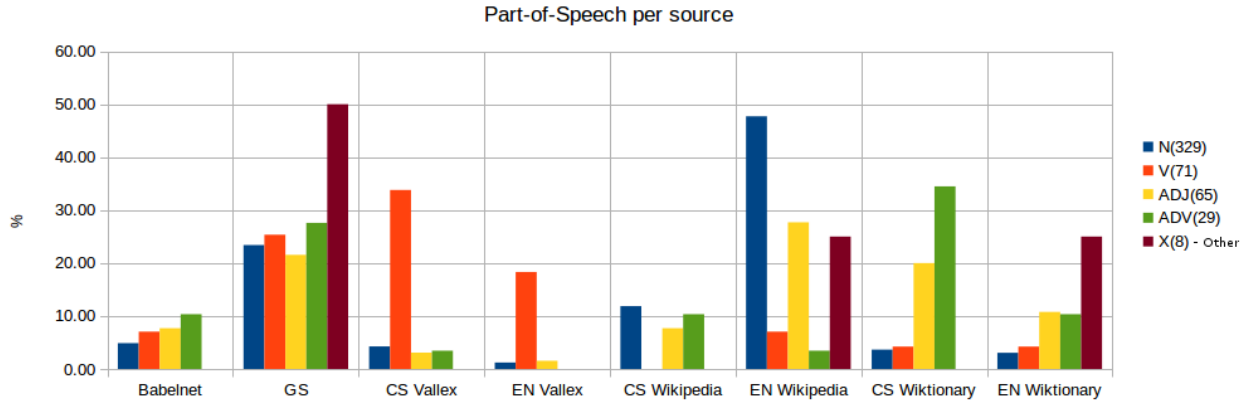
Vallex We scan the XML file and return all the frames belonging to the verb with the lemma matching the selected word's lemma.

4 First experiment

The first experiment was held in March 2014. The 7 participating annotators (none of whom had any experience in annotation tasks) were asked to annotate the sentences from PCEDT 2.0 [6] with Czech and English sources:

[6] http://ufal.mff.cuni.cz/pcedt2.0/en/index.html

Figure 2: Annotations from a given dictionary in the first experiment broken down by part of speech of the annotated words.



Wikipedia and Wiktionary for both languages, BabelNet, Google Search, and the Czech and English Vallexes. Each annotator was given a set of sentences in English or Czech and they were asked to annotate as many words or phrases in each sentence as possible, with as many reasonable meanings as they can. We required the annotators to annotate across parts of speech if possible (for instance to annotate the noun "teacher" with the corresponding verb "to teach"). This requirement appeared because we wanted to evaluate the possibility of using more abstract *senses* as used, for instance, in works with AMR.

4.1 Gathered annotations

In total, we collected 507 annotations for 158 units. 75 of these units had more than one annotation.

The upper part of Table 1 provides details on how often each of the annotation options was picked for a given source in the first annotation experiment. Note that in the first experiment, we did not offer the "Whole Page" option.

We see that the sources exhibit slightly different patterns of use. Wikipedia has lots of "Bad List" options selected due to the issue described in Section 4.2. GS is the most ambiguous resource, the user has picked two or more sense in about one half of GS annotations. The highest number of "Bad Lists" was received by the English Wikipedia (18 out of 40).

Figure 2 shows the distribution of different POS per source. Google Search seems to be the most versatile resource, covering all parts of speech well. The relatively low use of BabelNet was due to the web API usage limit. Vallexes work well for verbs but cross-POS annotation is only an exception. Wikipedia and Wiktionary are indeed somewhat complementary in covered POSes.

4.2 Bad List vs. None Issue

The "Bad List" annotations should be used in two cases: (1) when the system fails to extract the selection list from a good page, and (2) when the whole page is wrong, for example when the system shows the Wikipedia page "South Africa" for the word "south". "None" was meant for correct selection lists (matching domain, reasonable options) but the right option missing. The guidelines for the first experiment were not very clear on this so some annotators marked problems with selection list as "Bad List" and some used the label "None".

Manual revision revealed that only 10 out of 40 "Bad List" annotations were indeed "Bad List" in one of the two meanings described above. The right hand part of Table 2 shows IAA after changing wrongly annotated "Bad Lists" into "None".

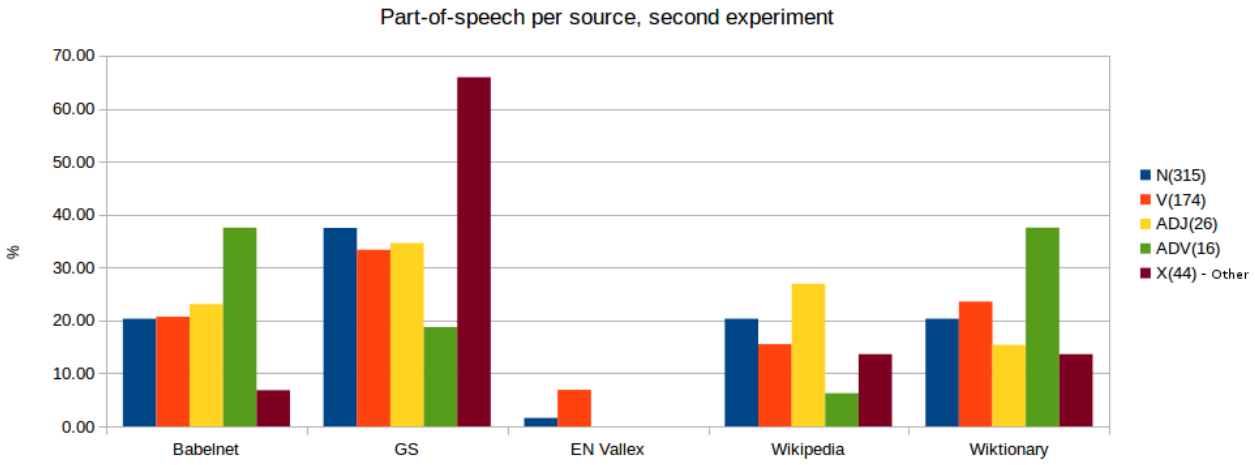
Table 2: Inter-annotator agreement in the first experiment, before (left) and after (right) the "Bad List" fix.

Source	Annotations	2-IAA	Annotations	2-IAA
Babelnet	29	0.69	29	0.69
GS	120	0.24	120	0.24
CS Vallex	46	0.58	46	0.58
EN Vallex	19	1.00	19	1.00
CS Wikipedia	47	0.32	43	0.35
EN Wikipedia	183	0.05	181	0.10
CS Wiktionary	38	0.29	38	0.35
EN Wiktionary	25	0	25	0
Total:	507	0.21	501	0.24

4.3 Inter-annotator agreement

Inter-annotator agreement is a measure of how well two annotators can make the same annotation decision for a certain item. In our case it is measured as the percentage of cases when a pair (2-IAA) of annotators agree on the (set of) senses for a given annotation unit. The measurement was made pairwise for all the annotations, which had more that one annotator. The results are presented in Table 2, before and after fixing the "Bad List" issue.

Figure 3: Annotations from a given dictionary in the second experiment broken down by part of speech of the annotated words.



In general, the IAA estimates should be treated with caution. Many units were assigned only to a single annotator, so they weren't taken into account while computing IAA.

The extremely low IAA for English Wikipedia was caused by the following issue. For several units, one annotator tried to select all the *senses* to show that the whole page can be used, while others have picked one or only a few *senses*. We resolved the issue by introducing a new option "Whole page" in the second experiment.

Interestingly, we see a negative correlation (Pearson correlation coefficient of -0.37) between the number of units annotated for a given source and the 2-IAA.

We also report Cohen's kappa [13], which reflects the agreement when disregarding agreement by chance. In our setting, we estimate the agreement by chance as one over the length of the selection list plus two (for "None" and "Bad List"). This is a conservative estimate, in principle the annotators were allowed to select any subset of selection list. We compute kappa using $K = \frac{P_a - P_e}{1 - P_e}$, where P_a was the total 2-IAA and P_e was the arithmetical average of agreements by chance for each annotation. Kappa for the first experiment was 0.13.

To assess the level of uncertainty for the estimates, we use bootstrap resampling with 1000 resamples, which gives us IAA of 0.25 ± 0.1 and kappa of 0.135 ± 0.115 for 95% of samples.

5 Second experiment

The second experiment was held in March 2015 with another group of 6 annotators. One of the annotators had experience in annotating tasks, while others had no such experience. The setting of the experiment was slightly different. The annotators were asked to annotate only English sentences from QTLeap project[7] using BabelNet, Google Search, English Wikipedia, English Wiktionary and ENGVALLEX. The guidelines were refined, asking the annotators to mark the largest possible span for each concept in the sentence, e.g. to annotate "mouse cursor" jointly as one concept and not separately as "computer pointing device" for the word "mouse" and "graphic representation of computer mouse on the screen" for the word "cursor". The option "Whole page" was newly introduced to help users indicate that the whole page can be used as a sense.

5.1 Gathered annotations

We collected 570 annotations for 35 words, 32 of which had annotations from more than one annotator. The number of units here is lower that in the first experiment, because all our annotators used the same sentences. Also, for the second experiment we required the annotators to use all the resources for each unit, so we have more results per unit.

During the second experiment, the system processed 147 unique (in terms of *selected word(s)* and *selected resource*) queries. All the resources got nearly equal number of queries (about 30), except for Vallex, which got only 10 queries. The annotators changed the queries 59 times, but this also includes cases, when Wikipedia used its own inner redirects, which our system did not distinguish from users' changes. BabelNet was changed 9 times, Google Search – 2, Vallex – 8, Wikipedia – 21 and Wiktionary – 19. Based on these numbers, GS may seem more reliable but it is not necessary true. One reason is that some of part of the changes for Wikipedia was made automatically by Wikipedia itself. The other argument is that users could limit their effort and after examining the first 10 GS re-

[7] http://qtleap.eu/

Table 3: Coverage per *content word* (second experiment). The left part reports the union across annotators, the right part reports the percentage of content words receiving a valid label (Labeled) for each annotator separately.

	Content words		Annotators					
Source	Attempted	Labeled	A1	A2	A3	A4	A5	A6
Babelnet	100%	**91%**	53%	20%	67%	66%	**79%**	40%
GS	100%	85%	50%	13%	53%	46%	**76%**	20%
Vallex	32%	26%	7%	6%	10%	**40%**	0%	0%
Wikipedia	100%	58%	39%	20%	35%	**53%**	50%	26%
Wiktionary	100%	88%	53%	20%	32%	40%	**76%**	26%
Total *content words*	34	34	28	15	28	15	34	15

Table 4: Inter-annotator agreement, second experiment

Source	Annotations number	2-IAA
Babelnet	114	0.49
GS	217	0.45
Vallex	17	0.60
Wikipedia	105	0.61
Wiktionary	117	0.28
Total:	570	0.46

sults for the query they just picked "Bad List" option and moved on, not trying to change the query.

The POS per source distribution (see Figure 3) for the second experiment is similar to the first one, except for the BabelNet, which did not reach any technical limit this time and was therefore used more often across all POSes.

5.2 Coverage

In Table 3, we show the coverage of *content words* in the second experiment. By *content words* we mean all the words in the sentence, except for auxiliary verbs, punctuation, articles and prepositions. The instructions asked to annotate all content words. Each annotator completed a different number of sentences, so the number of words annotated differs. The column *Content words Attempted* shows the total number of words with some annotation at all, while *Labeled* are words which received some sense, not just "None" or "Bad List". Both numbers are taken from the union over all annotators. Babelnet get the best coverage in terms of *Labeled* annotations. The right hand side of the table shows how many words each annotator has *labeled*. Since the union is considerably higher than the most productive annotator, we need to ask an important question: How many annotators do we need to have a perfect coverage of the sentence.

5.3 Inter-annotator agreement

Results presented in Table 4 are overall better than in the first experiment. The kappa was computed as in Section 4.3 with the only one difference: we added 3 instead of 2 options when estimating the local probability of the agreement by chance (for the new "Whole Page" option). Kappa for the second experiment was 0.40. Bootstrapping showed IAA 0.39 ± 0.055 and kappa 0.32 ± 0.06 for 95% central resamples. Again, the 2-IAA is negatively correlated with the number of units annotated (Pearson correlation coefficient -0.22).

6 Discussion

Comparing first and second experiment, one can see, that we managed to improve IAA by expanding the set of available options and refining the instructions, but IAA is still not satisfactory.

For resources where IAA reaches 60% (Vallex and Wikipedia), the coverage is rather low, 26% and 58%. BabelNet gives the best coverage but suffers in IAA. Google Search seems an interesting option for its versatility across parts of speech, on par with established knowledge bases like BabelNet in terms of inter-annotator agreement but with much more ambiguous "senses". The cross-POS annotation does not seem very effective in practice, but a more thorough analysis is desirable.

7 Comparison with Other Annotation Tools

Several automatic systems for sense annotation are available. Our dataset could be used to compare them empirically on the annotations from the respective repository used by each of the tools. For now we provide only an illustrative comparison of these three systems: TAGME[8], DBpedia Spotlight[9] ,and Babelfy[10]

Figure 4 provides an example of our manually collected annotations for the sentence "Move the mouse cursor to the beginning of the blank page and press the DELETE key as often as needed until the text is in the desired spot.".

For this sentence, the TAGME system with default settings returned three entities ("mouse cursor", "DELETE key" and "text"). DBpedia Spotlight with default settings (confidence level = 0.5) returned one entity ("mouse"). Babelfy showed the best result among these systems in terms of coverage, failing to recognize only the verb

[8] http://tagme.di.unipi.it/
[9] http://dbpedia-spotlight.github.io/demo/
[10] http://babelfy.org/

Figure 4: Our BabelNet and Wikipedia manual annotations and outputs of three automatic sense taggers for the sentence "Move the mouse cursor to the beginning of the blank page and press the DELETE key as often as needed until the text is in the desired spot." Overlap indicated by italics (BabelNet and Babelfy) and bold (Wikipedia and TAGME).

	BabelNet	Wikipedia	TAGME	Spotlight	Babelfy
Move	bn:00087012v,bn:00090948v bn:00056033n, bn:00056155n	Motion_(physics)	-	-	-
mouse	*bn:00021487n*,bn:00090942v *bn:00024529n*	mouse_(disambiguation), mouse_cursor	Mouse_(computing)	Mouse_(computing)	*bn:00024529n*,*bn:00021487n*
cursor	*bn:00024529n*	**mouse_cursor**, cursor_(disambiguation)	**mouse_cursor**	-	*bn:00024529n*
beginning	bn:00009632n,bn:00009633n bn:00009634n,bn:00009635n	beginning, beginning_(disambiguation)	-	-	bn:00083340v
blank	*bn:00098524a*	blank_page_(disambiguation)	-	-	bn:01161190n,*bn:00098524a*
page	*bn:00060158n*	blank_page_(disambiguation)	-	-	bn:01161190n,*bn:00060158n*
press	bn:00091988v,bn:00091986v	press_(disambiguation)	-	-	bn:00046094n
DELETE	*bn:01208543n*	**Delete_key**, DELETE	**Delete_key**	-	*bn:01208543n*, bn:00045088n
key	*bn:01208543n*, bn:00048996n	**Delete_key**, key_(disambiguation)	**Delete_key**	-	*bn:01208543n*, bn:00048985n
often	bn:00114048r, bn:00115452r bn:00116418r	often	-	-	-
needed	bn:00107194a	Need_(disambiguation)	-	-	bn:00082822v
until	-	until	-	-	-
text	*bn:00076732n*	text_(disambiguation)	Plain_text	-	*bn:00076732n*
desired	bn:00100580a, bn:00026550n bn:00100607a	Desire_(disambiguation), desired	-	-	bn:00086682v
spot	*bn:00062699n*	spot_(disambiguation)	-	-	*bn:00062699n*

"move" and adverbs "often" and "until", but it also provided several false meanings for found entities.

8 Conclusion

In this paper, we examined how different dictionaries can be used for entity linking and word sense disambiguation. In our unifying view based on finding the best "selection list" and selecting one or more senses from it, we tested standard inventories like BabelNet or Wikipedia, but also Google Search.

We proposed and refined annotation guidelines in two consecutive experiments, reaching average inter-annotator agreement of about 46%, with Wikipedia and Vallex up to 60%. Higher agreement seems to go together with lower coverage, but further investigation is needed for confirmation and to find the best balance of granularity, coverage and versatility among existing sources.

Acknowledgements

This research was supported by the grants FP7-ICT-2013-10-610516 (QTLeap). This research was partially supported by SVV project number 260 224. This work has been using language resources developed, stored and distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2010013).

References

[1] Demartini, G., et al.: Zencrowd: leveraging probabilistic reasoning and crowdsourcing techniques for large-scale entity linking. In: Proceedings of the 21st international conference on World Wide Web, ACM (2012) 469–478

[2] Bennett, P.N., et al.: Report on the sixth workshop on exploiting semantic annotations in information retrieval (ESAIR'13). In: ACM SIGIR Forum. Volume 48., ACM (2014) 13–20

[3] Ratinov, L., et al.: Local and global algorithms for disambiguation to wikipedia. In: Proc. of ACL/HLT, Volume 1. (2011) 1375–1384

[4] Navigli, R.: Word sense disambiguation: A survey. ACM Comput. Surv. **41**(2) (February 2009) 10:1–10:69

[5] Pereira, B.: Entity linking with multiple knowledge bases: An ontology modularization approach. In: The Semantic Web–ISWC 2014. Springer (2014) 513–520

[6] Moro, A., Navigli, R.: SemEval-2015 Task 13: Multilingual All-Words Sense Disambiguation and Entity Linking. In: Proc. of SemEval-2015. (2015) In press.

[7] Ferragina, P., Scaiella, U.: Tagme: on-the-fly annotation of short text fragments (by wikipedia entities). In: Proc. of CIKM, ACM (2010) 1625–1628

[8] Zhang, L., Rettinger, A., Färber, M., Tadić, M.: A comparative evaluation of cross-lingual text annotation techniques. In: Information Access Evaluation. Multilinguality, Multimodality, and Visualization. Springer (2013) 124–135

[9] Banarescu, L., , et al.: Abstract Meaning Representation for Sembanking (2013)

[10] Navigli, R., Ponzetto, S.P.: BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artificial Intelligence **193** (2012) 217–250

[11] Žabokrtský, Z., Lopatková, M.: Valency information in VALLEX 2.0: Logical structure of the lexicon. The Prague Bulletin of Mathematical Linguistics (87) (2007) 41–60

[12] Lopatková, M., Žabokrtský, Z., Ketnerová, V.: Valenční slovník českých sloves. (2008)

[13] Cohen, J.: A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement **20**(1) (1960)