Linked Disambiguated Distributional Semantic Networks

Stefano Faralli[1], Alexander Panchenko[2], Chris Biemann[2], and Simone P. Ponzetto[1]

[1] Data and Web Science Group, University of Mannheim, Germany
{stefano,simone}@informatik.uni-mannheim.de,
[2] Language Technology Group, TU Darmstadt, Germany
{panchenko,biem}@lt.informatik.tu-darmstadt.de,

Abstract. We present a new hybrid lexical knowledge base that combines the contextual information of distributional models with the conciseness and precision of manually constructed lexical networks. The computation of our count-based distributional model includes the induction of word senses for single-word and multi-word terms, the disambiguation of word similarity lists, taxonomic relations extracted by patterns and context clues for disambiguation in context. In contrast to dense vector representations, our resource is human readable and interpretable, and thus can be easily embedded within the Semantic Web ecosystem.

Resource type: Lexical Knowledge Base
Permanent URL: https://madata.bib.uni-mannheim.de/id/eprint/171

1 Introduction

Recent years have witnessed an impressive amount of work on the automatic construction of wide-coverage knowledge resources from Wikipedia [3, 13] and the Web [7]. Complementary to this, a plethora of works in Natural Language Processing (NLP) has recently focused on combining knowledge bases with distributional information from text. These include approaches that modify Word2Vec [15] to learn sense embeddings [5], methods to enrich WordNet with embeddings for synsets and lexemes [21], acquire continuous word representations by combining random walks over knowledge bases and neural language models [11], or produce joint lexical and semantic vectors for sense representation from text and knowledge bases [4]

In this paper, we follow this line of research and take it one step forward by producing a hybrid knowledge resource, which combines symbolic and statistical meaning representations while i) staying purely on the lexical-symbolic level, ii) explicitly distinguishing word senses, and iii) being human readable. Far from being technicalities, such properties are crucial to be able to embed a resource of this kind into the Semantic Web ecosystem, where human-readable distributional representations are explicitly linked to URIfied semantic resources. To this end, we develop a methodology to automatically induce distributionally-based semantic representations from large amounts of text, and link them to a reference knowledge base. This results in a new knowledge resource that we refer to as a *hybrid aligned resource*.

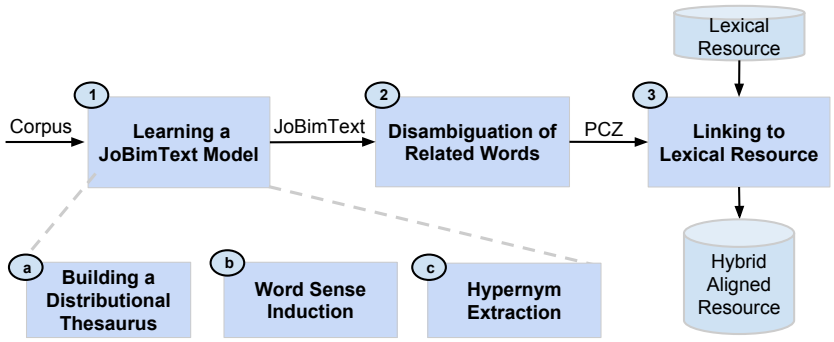

Fig. 1: Overview of our method for constructing a hybrid aligned resource.

2 Building a Hybrid Aligned Resource

Our resource is built in three main phases (Figure 1):

1) **Learning a JoBimText model**: initially, we automatically create a sense inventory from a large text collection using the pipeline of the JoBimText project [2, 22][1]. The resulting structure contains disambiguated proto-concepts (i.e. senses), their similar and related terms, as well as aggregated contextual clues per proto-concept (Table 1a). This is a distributionally-based conceptualization with some degree of taxonomic information only. Hence the two subsequent phases, which, together with the final resource, represent the novel contribution of this paper.
2) **Disambiguation of related terms**: we fully disambiguate all lexical information associated with a proto-concept (i.e. similar terms and hypernyms) based on the partial disambiguation from step 1). The result is a proto-conceptualization (PCZ). In contrast to a term-based distributional thesaurus (DT), a PCZ consists of sense-disambiguated entries, i.e. all terms have a sense identifier (Table 1b).
3) **Linking to a lexical resource**: we align the PCZ with an existing lexical resource (LR). That is, we create a mapping between the two sense inventories and then combine them into a new extended sense inventory, our *hybrid aligned resource*.

2.1 Learning a JoBimText model

Following [2], we apply a holing operation where each observation in the text is split into a term and its context. The 1000 most significant contexts per term, as determined by the LMI significance measure [8], serve as a representation for the term, and term similarity is defined as the number of common contexts. This procedure induces a weighted similarity graph over terms, also known as Distributional Thesaurus (DT), where each entry of the DT consists of the most similar 200 terms for a given term.

In DTs, entries of polysemous terms t are mixed, i.e. they contain similar terms stemming from several senses respectively usages of the term. Since terms that belong

[1] `http://www.jobimtext.org`

entry	similar terms	hypernyms	context clues
mouse:NN:0	rat:NN, rodent:NN, monkey:NN, ...	animal:NN, species:NN, ...	rat:NN:conj_and, white-footed:JJ:amod, ...
mouse:NN:1	keyboard:NN, computer:NN, printer:NN ...	device:NN, equipment:NN, ...	click:NN:-prep_of, click:NN:-nn, ...
keyboard:NN:0	piano:NN, synthesizer:NN, organ:NN ...	instrument:NN, device:NN, ...	play:VB:-dobj, electric:JJ:amod, ..
keyboard:NN:1	keypad:NN, mouse:NN, screen:NN ...	device:NN, technology:NN ...	computer:NN:nn, qwerty:JJ:amod ...

(a) JoBimText model entries

entry	similar terms	hypernyms	context clues
mouse:NN:0	rat:NN**:0**, rodent:NN**:0**, monkey:NN**:0**, ...	animal:NN**:0**, species:NN**:1**, ...	rat::NN:conj_and, white-footed:JJ:amod, ...
mouse:NN:1	keyboard:NN**:1**, computer:NN**:0**, printer:NN**:0** ...	device:NN**:1**, equipment:NN**:3**, ...	click:NN:-prep_of, click:NN:-nn, ...
keyboard:NN:0	piano:NN**:1**, synthesizer:NN**:2**, organ:NN**:0** ...	instrument:NN**:2**, device:NN**:3**, ...	play:VB:-dobj, electric:JJ:amod, ..
keyboard:NN:1	keypad:NN**:0**, mouse:NN**:1**, screen:NN**:1** ...	device:NN**:1**, technology:NN**:0** ...	computer:NN:nn, qwerty:JJ:amod ...

(b) Proto-conceptualization entries

Table 1: Examples of entries for "mouse" and "keyboard" from the *news-p1.6* dataset before and after the semantic closure. Trailing numbers indicate sense identifiers.

to the same sense are more similar to each other than to terms belonging to a different sense, we can employ graph clustering to partition the open neighbourhood of t in the DT (i.e., terms similar to t and their similarities, without t) to arrive at sense representations for t, characterized by a list of similar terms. We achieve this by applying Chinese Whispers [1] on the ego-network of the term t, as defined by its similar terms as nodes.

Further, we run Hearst patterns [12] over the corpus to extract IS-A (hypernym) relations between terms. We add these hypernyms to senses by aggregating IS-A relations over the list of similar terms for the given sense into a weighted list of hypernyms. Additionally, we aggregate the significant contexts of similar terms per sense to arrive at weighted aggregated context clues. The resulting structure is called the JoBimText model [22] of the corpus. A JoBimText entry consists of a distributionally-induced word sense, a ranked list of similar terms for this sense, a list of superordinate terms and a list of aggregated context clues (note that only unstructured text is required). Table 1(a) shows some JoBimText entries for the polysemous terms "mouse" and "keyboard".

2.2 Disambiguation of related terms

While JoBimText models contain sense distinctions, they are not fully disambiguated: the list of similar and hypernyms terms of each sense does not carry sense information. In our example (Table 1a) the sense of "mouse:NN" for the entry "keyboard:NN:1" could either be the "animal" or the "electronic device" one. Consequently, we next apply a semantic closure procedure to arrive at a PCZ in which *all* terms get assigned a unique, best-fitting sense identifier (Table 1b).

At its heart, our method assigns each target word w to disambiguate – namely, a similar and superordinate term from each sense of the JoBimText model – the sense $\hat{s}$ whose context (i.e., the list of of similar or superordinate terms) has the maximal similarity with the target word's context (i.e., the other words in the target word's list of similar or superordinate items) – we use cosine as similarity metric:

$$\hat{s} = \underset{s \in Senses_{JoBim}(w)}{\operatorname{argmax}} \; cos(ctx(w), ctx(s)). \tag{1}$$

This way we are able to link, for instance, *keyboard:NN* in the list of similar terms for *mouse:NN:1* to its 'device' sense (*keyboard:NN:1*), since *mouse:NN:1* and *keyboard:NN:1* share a large amount of terms from the IT domain.

The structure of a PCZ resembles that of a lexical semantic resource: each term has a list of proto-concepts, and proto-concepts are linked via relations, such as similarity and taxonomic links. Sense distinctions and distributions are dependent on the underlying corpus, which causes the PCZ to naturally adapt to the domain of the corpus. A large difference to manually created resources, however, is the availability of aggregated context clues that allow to disambiguate polysemous terms in text with respect to their sense distinctions. Table 1(b) shows example proto-concepts for the terms "mouse:NN" and "keyboard:NN", taken from our *news-p1.6* PCZ (see Section 3.1).

2.3 Linking to a lexical resource

We next link each sense in our proto-conceptualization (PCZ) to the most suitable sense (if any) of a Lexical Resource (LR, see Figure 1 step 3). Our method takes as input:

1. a PCZ $T = \{(j_i, R_{j_i}, H_{j_i})\}$ where j_i is a sense identifier (i.e. *mouse:NN:1*), R_{j_i} the set of its semantically related senses (i.e. $R_{j_i} = \{$*keyboard:NN:1*, *computer:NN:0*, $\dots\}$ and H_{j_i} the set of its hypernym senses (i.e. $H_{j_i} = \{$*equipment:NN:3*, $\dots\}$;
2. a LR W: we experiment with: WordNet [10], a lexical database for English and BabelNet [16], a very large multilingual "encyclopaedic dictionary";
3. a threshold th over the similarity between pairs of concepts and a number m of iterations as stopping criterion;

and outputs a mapping M, which consists of a set of pairs of the kind $(source, target)$ where $source \in T.senses$ is a sense of the input PCZ T and $target \in W.senses \cup source$ is the most suitable sense of W or $source$ when no such sense is available. At its heart, the mapping algorithm compares the senses across resources with the following similarity function:

$$sim(j, c, M) = \frac{|T.BoW(j, M, W) \cap W.BoW(c)|}{|T.BoW(j, M, W)|}, \quad \text{where:} \tag{2}$$

1. $T.BoW(j, M, W)$ is the set of words containing all the terms extracted from related/hypernym senses of j and all the terms extracted from the related/hypernym (i.e., already linked in M) synsets in W. For each synset we use all synonyms and content words of the gloss.
2. $W.BoW(c)$ contains the synonyms and the gloss content words for the synset c and all the related synsets of c.

A new link pair (j, c) is then added to M if the similarity score between j and c is greater than or equal to a threshold th. Finally, all unlinked j of T, i.e. proto-concepts that have no corresponding LR sense, are added to the mapping M. We follow the guidelines from McCrae et al. [14] and create an RDF representation to share the mapping between our PCZs and lexical knowledge graphs (i.e., WordNet and BabelNet) in the Linked Open Data Cloud [6].

		words			senses	polysemy		rel. senses		hypernyms	
dataset	n	#	monosemous	polysemous	#	avg.	max	#	avg.	#	avg.
news-p1.6	200	207k	137k	69k	332k	1.6	18	234k	63.9	15k	6.9
news-p2.3	50	200k	99k	101k	461k	2.3	17	298k	44.3	15k	5.8
wiki-p1.8	200	206k	120k	86k	368k	1.8	15	300k	59.3	15k	4.4
wiki-p6.0	30	258k	44k	213k	1.5M	6.0	36	811k	16.9	52k	1.7
wiki-mw-p1.6	200	465k	288k	176k	765k	1.6	13	662k	46.6	30k	3.2

Table 2: Structural analysis of our five proto-conceptualizations (PCZs).

3 Experiments

3.1 Datasets for the extraction of the proto-conceptualizations (PCZs)

We experiment with two different large corpora, namely a 100 million sentence news corpus (*news*) from Gigaword [17] and LCC [19], and with a 35 million sentence Wikipedia corpus (*wiki*) and different parametrizations of the sense induction algorithm to obtain five proto-conceptualizations (PCZ) with different average sense granularities. Further, we use the method described in [20] to compute a dataset that includes automatically extracted multiword terms. In Table 2, we present figures for our five datasets. For each dataset, Columns 3, 4 and 5 count the overall number of words, including monosemous words and polysemous ones, respectively. For each PCZ we report the cardinality (Column 6), the average polysemy (Column 7) and the maximum polysemy (Column 8). Finally, we report the overall and the average number of related senses and hypernyms (Column 9-12).

3.2 Experiment 1: disambiguation of the distributional thesaurus entries

Experimental setting. In order to disambiguate a related or superordinate term t in a word sense entry s in the JoBimText model, we compare the related words of s with the related words of each of the senses t_s for the target term t. Similarly, we evaluate the quality of the disambiguation of the JoBimText models by judging the compatibility of the similar words for s and t_s. For instance, the similar term *mouse:NN*, in the JoBimText model entry for keyboard:NN:1, namely "keypad:NN, *mouse:NN*, screen:NN, ..." is compatible with the related words "keyboard:NN, computer:NN, printer:NN ..." (i.e., those of sense mouse:NN:1) and is not compatible with the related words "cat:NN, rodent:NN, monkey:NN, ..." of mouse:NN:0 (see Table 1).

Our experimental setting is based on three steps: 1) we manually select 17 highly ambiguous *target words*; 2) we collect 19,774 disambiguated entries of the *wiki-p1.6* JoBimText model where the target words appear and randomly sample 15% of these entries to make annotation feasible; 3) we manually judge entries in the sample on whether the related words of the target word fits the sense assigned or not. [2]. Finally, we compute performance by means of standard accuracy – i.e., the proportion of cases in which the similar or hypernym terms from the JoBimText model are correctly disambiguated.

[2] The target words and annotations can be found at `https://goo.gl/jjdhI4`.

Results and discussion. Our method achieves an accuracy of 0.84 across all parts of speech, including accuracy of 0.94 for nouns, 0.85 for proper nouns, 0.76 for verbs, and 0.63 for adjectives. Different results across parts of speech are due to the different quality of the respective DT clusters. This is because this first experiment also indirectly measures the quality of the senses from the JoBimText model: indeed, a match between two word sense entries is only possible if both of them are interpretable.

To better understand the amount of spurious items in our sense clusters, we performed an additional manual evaluation where, for a sample of 100 randomly sampled noun PCZ items, we counted the ratio between wrong (e.g., *rat* for the computer sense of *mouse*) and correct (*keyboard*, *computer*, etc.) related concepts that were found within the PCZs. We obtained a macro average of 0.0495 and a micro average of 0.0385 wrong related concepts within the PCZs. Moreover, 83% of the above sample has no unrelated concepts, and only 2% has only one unrelated concept with a macro average ratio between the wrong and corrects related PCZ of 0.067. This indicates that, overall, the amount of spurious concepts within clusters is indeed small, thus providing a high-quality context for an accurate disambiguation of noun DT clusters.

3.3 Experiment 2: linking to lexical knowledge bases

Experimental setting. Next, we evaluate the performance of our linking component (Section 2.3). For this, we choose two lexical-semantic networks: WordNet [10], which has a high coverage on English common nouns, verbs and adjectives, and BabelNet [16], which also includes a large amount of proper nouns as well as senses gathered from multiple other sources, including Wikipedia.

We follow standard practices (e.g., [16]) and create five evaluation test sets, one for each dataset from Section 3.1, by randomly selecting a subset of 300 proto-concepts for each dataset, and manually establishing a mapping from these senses to WordNet and BabelNet concepts (proto-concepts that cannot be mapped are labeled as such in the gold standard). The quality and correctness of the mapping is estimated as accuracy on the ground-truth judgments, namely the amount of true mapping decisions among the total number of (potentially, empty) mappings in the gold standard. We also evaluate our mapping by quantifying Coverage (the percentage of senses of the knowledge base sense inventory covered by the mapping M) and ExtraCoverage (the ratio of concepts in M not linked to the knowledge base sense inventory over the total number of knowledge base senses). The latter is a measure of novelty to quantify the amount of senses discovered in T and not represented by the knowledge base: it indicates the amount of 'added' knowledge we gain with our resource based on the amount of proto-concepts that cannot be mapped and are thus included as novel senses.

Results and discussion. In Table 3 we present the results using the optimal parameter values (i.e. th=0.0 and m=5)[3]. For all datasets the number of linked senses, Coverage and ExtraCoverage are directly proportional to the number of entries in the dataset –

[3] To find optimal value for m, we prototyped our approach on a dev set consisting of a random sample of 300 proto-concepts, and studied the curves for the number of linked proto-concepts to WordNet resp. BabelNet. The th value was then selected as to maximize the accuracy.

	WordNet				BabelNet			
dataset	⚓ senses	Cov.	ExtraCov.	Accuracy	⚓ senses	Cov.	ExtraCov.	Accuracy
news-p1.6	88k	34.5%	206.0%	86.9%	164k	1.3%	2.9%	81.8%
news-p2.3	145k	38.2%	267.0%	93.3%	236k	1.4%	3.9%	85.1%
wiki-p1.8	91k	35.9%	234.7%	94.8%	232k	1.9%	2.4%	86.4%
wiki-p6.0	400k	49.9%	919.9%	93.5%	737k	2.8%	1.3%	82.2%
wiki-mw-p1.6	81k	30.7%	581.2%	95.3%	589k	4.7%	1.8%	83.8%

Table 3: Results on linking to lexical knowledge bases: number of linked proto-concepts (⚓), Coverage, ExtraCoverage and Accuracy for our five datasets.

i.e., the finer the concept granularity, as given by a lower sense clustering n parameter, the lower the number of mapped senses, Coverage and ExtraCoverage.

In general, we report rather low coverage figures: the coverage in WordNet is always lower than 50% (30% in one setting) and in BabelNet is in all settings lower than 5%. Low coverage is due to different levels of granularities between the source and target resource. Our target knowledge bases, in fact, have very fine-grained sense inventories. For instance, BabelNet lists 17 senses of the word "python" including two (arguably obscure ones) referring to particular roller coasters. In contrast, word senses induced from text corpora tend to be coarse and corpus-specific. Consequently, the low coverage comes from the fact that we connect a coarse and a fine-grained sense inventory – cf. also previous work [9] showing comparable proportions between coverage and extra-coverage of automatically acquired knowledge (i.e., glosses) from corpora.

Finally, our results indicate differences between the order of magnitude of the Coverage and ExtraCoverage when linking to WordNet and BabelNet. This high difference depends on the cardinality of the two sense inventories, where BabelNet has millions of senses while WordNet more than a hundred of thousands – many of them not covered in our corpora. Please note that an ExtraCoverage of about 3% in BabelNet corresponds to about 300k novel senses. Overall, we take our results to be promising in that, despite the relative simplicity of our approach (i.e., almost parameter-free unsupervised linking), we are able to reach high accuracy figures in the range of around $87\% - 95\%$ for WordNet and accuracies above 80% for BabelNet.

4 Conclusions

We presented an automatically-constructed hybrid aligned resource that combines distributional semantic representations with lexical knowledge graphs. To the best of our knowledge, we are the first to present such a large-scale, fully URIfied hybrid aligned resource with high alignment quality. As future work, we will explore ways to couple focused crawling [18] with domain-specific PCZs to extend our resource to many domains. Moreover, we aim at using our linguistically-grounded hybrid resource to provide generalizations beyond concepts, such as, for instance, hybrid symbolic and distributional representations of actions and events.

Acknowledgments. We acknowledge the support of the Deutsche Forschungsgemeinschaft (DFG) under the JOIN-T project.

References

1. Biemann, C.: Chinese Whispers: An efficient graph clustering algorithm and its application to natural language processing problems. In: Proc. TextGraphs. pp. 73–80 (2006)
2. Biemann, C., Riedl, M.: Text: Now in 2D! A Framework for Lexical Expansion with Contextual Similarity. Journal of Language Modelling 1(1), 55–95 (2013)
3. Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., Hellmann, S.: DBpedia – A crystallization point for the web of data. JWS 7(3), 154–165 (2009)
4. Camacho-Collados, J., Pilehvar, M.T., Navigli, R.: NASARI: a novel approach to a semantically-aware representation of items. In: Proc. NAACL-HLT. pp. 567–577 (2015)
5. Chen, X., Liu, Z., Sun, M.: A unified model for word sense representation and disambiguation. In: Proc. EMNLP. pp. 1025–1035 (2014)
6. Chiarcos, C., Hellmann, S., Nordhoff, S.: Linking linguistic resources: Examples from the Open Linguistics Working Group. In: Chiarcos, C., Nordhoff, S., Hellmann, S. (eds.) Linked Data in Linguistics - Representing and Connecting Language Data and Language Metadata., pp. 201–216. Springer (2012)
7. Dong, X., Gabrilovich, E., Heitz, G., Horn, W., Lao, N., Murphy, K., Strohmann, T., Sun, S., Zhang, W.: Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In: Proc. KDD. pp. 601–610 (2014)
8. Evert, S.: The Statistics of Word Cooccurrences: Word Pairs and Collocations. Ph.D. thesis, Institut für maschinelle Sprachverarbeitung, University of Stuttgart (2005)
9. Faralli, S., Navigli, R.: Growing multi-domain glossaries from a few seeds using probabilistic topic models. In: Proc. EMNLP. pp. 170–181 (2013)
10. Fellbaum, C. (ed.): WordNet: An Electronic Database. MIT Press, Cambridge, MA (1998)
11. Goikoetxea, J., Soroa, A., Agirre, E.: Random walks and neural network language models on knowledge bases. In: Proc. NAACL HLT. pp. 1434–1439 (2015)
12. Hearst, M.A.: Automatic acquisition of hyponyms from large text corpora. In: Proc. COLING. pp. 539–545 (1992)
13. Hoffart, J., Suchanek, F.M., Berberich, K., Weikum, G.: YAGO2: a spatially and temporally enhanced knowledge base from Wikipedia. ArtInt. pp. 28–61 (2013)
14. McCrae, J.P., Fellbaum, C., Cimiano, P.: Publishing and Linking WordNet using Lemon and RDF. In: Proceedings of the 3rd Workshop on Linked Data in Linguistics (2014)
15. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Proc. NIPS. pp. 3111–3119 (2013)
16. Navigli, R., Ponzetto, S.P.: BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. ArtInt. 193, 217–250 (2012)
17. Parker, R., Graff, D., Kong, J., Chen, K., Maeda, K.: English Gigaword Fifth Edition. Linguistic Data Consortium, Philadelphia (2011)
18. Remus, S., Biemann, C.: Domain-specific corpus expansion with focused webcrawling. In: Proc. LREC (2016)
19. Richter, M., Quasthoff, U., Hallsteinsdóttir, E., Biemann, C.: Exploiting the Leipzig corpora collection. In: Proc. IS-LTC (2006)
20. Riedl, M., Biemann, C.: A single word is not enough: Ranking multiword expressions using distributional semantics. In: Proc. EMNLP. pp. 2430–2440 (2015)
21. Rothe, S., Schütze, H.: AutoExtend: Extending word embeddings to embeddings for synsets and lexemes. In: Proc. ACL. pp. 1793–1803 (2015)
22. Ruppert, E., Kaufmann, M., Riedl, M., Biemann, C.: JoBimViz: A web-based visualization for graph-based distributional semantic models. In: Proc. ACL-IJCNLP System Demonstrations. pp. 103–108 (2015)