A Contrastive Evaluation of Word Sense Disambiguation Systems for Finnish

Frankie Robertson
University of Jyväskylä
Faculty of Information Technology
frankie.r.robertson@student.jyu.fi

Abstract

Previous work in Word Sense Disambiguation (WSD), like many tasks in natural language processing, has been predominantly focused on English. While there has been some work on other languages, including Uralic languages, up until this point no work has been published providing a contrastive evaluation of WSD for Finnish, despite the requisite lexical resources, most notably FinnWordNet, having long been in place. This work rectifies the situation. It gives results for systems representing the major approaches to WSD, including some of the systems which have performed best at the task for English. It is hoped these results can act as a baseline for future systems, including both multilingual systems and systems specifically targeting Finnish, as well as point to directions for other Uralic languages.

Tiivistelmä

Aiempi saneiden alamerkitysten yksiselitteistämistä käsittelevä työ, kuten monet muut luonnollisen kielen käsittelyyn liittyvät tehtävät, on enimmäkseen keskittynyt englannin kieleen. Vaikka hieman työtä on tehty myös muilla kielillä, mukaan lukien uralilaiset kielet, vertailevaa arviointia suomen kielen saneiden alamerkitysten yksiselitteistämisestä ei ole tähän mennessä julkaistu huolimatta siitä, että tarvittavat leksikaaliset resurssit, erityisesti FinnWordNet, ovat jo pitkään olleet saatavilla. Tämä työ pyrkii korjaamaan tilanteen. Se tarjoaa tuloksia merkittävimpiä lähestymistapoja saneiden alamerkitysten yksiselitteistämiseen edustavista ohjelmista, sisältäen joitakin parhaiten englanninkielellä samasta tehtävästä suoriutuvia ohjelmia. Näiden tulosten toivotaan voivan toimia lähtökohtana tuleville, sekä monikielisille että erityisesti suomen kieleen kohdentuville, ohjelmille ja tarjota suuntaviivoja muihin uralilaisiin kieliin keskittyvään työhön.

1 Introduction

Like many natural language understanding tasks, Word Sense Disambiguation (WSD) has been referred to as AI-complete (Mallery, 1988, p. 57). That is to say, it is considered as hard as the central problems in artificial intelligence, such as passing the

Proceedings of the fifth Workshop on Computational Linguistics for Uralic Languages, pages 42–54,
Tartu, Estonia, January 7 - January 8, 2019.

Turing test (Turing, 1950). While in the general case this may be true, the best current systems can at least do better than the (quite tough to beat) Most Frequent Sense (MFS) baseline. Evaluations against common datasets and dictionaries, largely following procedures set out by the shared tasks under the auspices of the SensEval and SemEval workshops, have been key to creating measurable progress in WSD.

For English, Raganato et al. (2017) present a recent comparison of different WSD systems across harmonised SensEval and SemEval data sets. Within the Uralic languages, Kahusk et al. (2001) created a manually sense annotated corpus of Estonian so that it could be included in SensEval-2. Two systems based on supervised learning were submitted, presented by Yarowsky et al. (2001) and Vider and Kaljurand (2001). Both systems failed to beat the MFS baseline (Edmonds, 2002, Table 1). For Hungarian, Mihálta (2010) created a sense tagged corpus by translating sense tagged data from English into Hungarian and then performed WSD with a number of supervised systems. Precision was compared with an MFS baseline, but the comparison was only given on a per-word basis. Up until this point, however, no work providing this type of a contrastive evaluation of WSD has been published for Finnish. This work rectifies the situation, giving results for systems representing the major approaches to WSD, including some of the systems which have performed best at the task for other languages.

2 Data and Resources

The minimum resources required to conduct a WSD evaluation are a Lexical Knowledge Base (LKB) and an evaluation corpus. Supervised systems require additionally a training corpus. The current generation of NLP systems make copious usage of word embeddings as lexical resources, as do some of the systems evaluated here, and so these are also needed. Here, the FinnWordNet (FiWN) (Lindén and Carlson, 2010) LKB is used, while both the evaluation and training corpus are based on the EuroSense corpus (Bovi et al., 2017). The rest of this section describes these linguistic resources and their preparation in more depth.

2.1 Obtaining a Sense Tagged Corpus

EuroSense (Bovi et al., 2017) is a multilingual sense tagged corpus, obtained by running the knowledge based Babelfy (Moro et al., 2014) WSD algorithm on multilingual texts. To use this corpus in a way which is compatible with the maximum number of systems and in line with the standards of previous evaluations, it first has to be preprocessed. The preprocessing pipeline is shown in Figure 1.

In the first stage, *drop non-Finnish*, all non Finnish text and annotations are removed from the stream. EuroSense is tagged with synsets from the BabelNet LKB (Navigli and Ponzetto, 2012). This knowledge base is based on the WordNets of many languages enriched and modified according to other sources, such as Wikipedia and Wikitionary. However, here the LKB to be used is FinnWordNet. A mapping file was extracted from BabelNet using its Java API and a local copy, obtained through direct communication with its authors[1]. The *Babelnet lookup* stage applies this mapping. The stage will drop annotation which do not exist in FiWN according to the mapping. A BabelNet synset can also map to multiple FiWN synsets, and in this case an ambiguous annotation can be produced.

[1]Made available at `https://github.com/frankier/babelnet-lookup`.

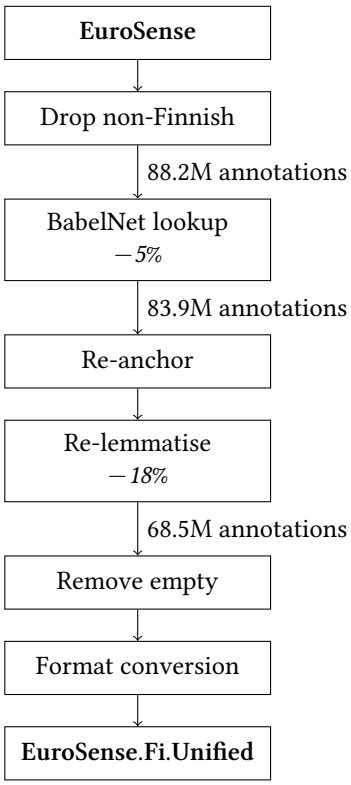

Figure 1: A diagram showing the pipeline to convert EuroSense to the unified format used for training and evaluation data. The number of annotations after various pipeline stages in millions are given, as are the proportion of annotations dropped by individual pipeline stages. The total proportion of Finnish annotations dropped is 22%.

The *re-anchor* and *re-lemmatise* stages clean up some problems with the grammatical analyses in EuroSense. EuroSense anchors sometimes include help words associated with certain verb conjugations, for example negative forms, e.g. "ei mene", or the perfect construction "on käynyt". *Re-anchor* removes these words from the anchor, taking care of the cases in which the whole anchor could actually refer to a lemma form in WordNet, e.g. "olla merkitystä". *Re-lemmatise* checks that the current lemma is associated with the annotated synsets in FiWN. In case there is no matching synsets, we look back at the surface form and check all possible lemmas obtained from OMorFi (Pirinen, 2015)[2] for matches against FiWN. At this point, any annotations which do not have exactly one lemma and one synset which exist in FiWN are dropped. In the penultimate stage, *remove empty*, any sentences without any annotations are removed entirely. Finally, the XML format is converted from the back-off annotations of the EuroSense format to the inline annotations of the unified format of Raganato et al. (2017).

The corpus is then split into testing and training sections. The testing corpus is made up of the first 1000 sentences, resulting in 4507 tagged instances. The resulting corpus is already sentence and word segmented. Additionally, the instance to be disambiguated is passed to each system with the correct lemma and part of speech tag, meaning the evaluation only tests the disambiguation stage of a full WSD pipeline and not the candidate extraction or POS tagging stage. The corpus is further processed with FinnPOS (Silfverberg et al., 2016)[3] for systems that need POS tags and/or lemmas for the words in the context.

2.2 Enriching FinnWordNet with frequency data

Many WSD techniques based on WordNet, including the typical implementation of the MFS baseline, assume it is possible to pick the most frequent sense of a lemma by picking the first sense. The reason this works with Princeton WordNet (PWN) (Miller et al., 1990) is because word senses are numbered according to the descending order of sense occurrence counts based on the part of the Brown corpus used during its creation[4]. FinnWordNet senses on the other hand are randomly ordered.

Since this data is potentially needed even by knowledge based systems, which should not have access to a training corpus, it is estimated here based on the frequency data in PWN. Unlike most PWN aligned WordNets, which are aligned at the synset level, FinnWordNet is aligned with PWN at the lemma level. An example of when this distinction takes effect is when lemmas are structurally similar. For example, in the synset "singer, vocalist, vocalizer, vocaliser", the Finnish lemma laulaja is mapped only to singer rather than to every lemma in the synset. When there is no clear distinction to be made, whole synsets are mapped. This reasoning fits with the existing structure of PWN: Relations between synsets encode purely semantic concerns, whereas relations between lemmas encode so-called morpho-semantic relationships, such as morphological derivation.

Let the Finnish-English lemma mapping be denoted $\mathcal{L}$, the specific frequency estimate for a Finnish lemma is then defined like so:

$$\text{freq}(l_{\text{fin}}) = \sum_{(l_{\text{fin}},\, l_{\text{eng}}) \in \mathcal{L}} \frac{\text{freq}(l_{\text{eng}})}{\left|\left\{(l_{\text{fin}_2},\, l_{\text{eng}}) \in \mathcal{L}\right\}\right|}$$

[2] `https://github.com/flammie/omorfi`

[3] `https://github.com/mpsilfve/FinnPos`

[4] This data is overlapping with, but distinct from SemCor (Miller et al., 1993).

Table 1: Word embeddings used

Name	Training data	Dim	Represents	Subword	Cross-lingual
MUSE Supervised fastText[ab]	Wikipedia & bilingual dictionary	300	Word forms	Yes	Yes
ConceptNet Numberbatch 17.06[cd]	Wikipedia & ConceptNet	300	Lemmas & Multiwords	—	Yes
NLPL Word2Vec[ef]	Wikipedia & CommonCrawl[g]	100	Word forms	No	No

[a] Conneau et al. (2017)
[b] `https://github.com/facebookresearch/MUSE`
[c] Speer et al. (2016)
[d] `https://github.com/commonsense/conceptnet-numberbatch`
[e] Fares et al. (2017)
[f] `http://vectors.nlpl.eu/repository/`
[g] `https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-1989`

The rationale of this approach is that this causes the frequencies of English lemmas to be evenly distributed across all the Finnish lemmas which they map to.

To integrate the resulting synthetic frequency data into as many applications as possible, it is made available in the WordNet format[5]. The WordNet format requires sense occurrence counts, meaning the frequency data must be converted to integer values. To perform this conversion all frequencies are multiplied by the lowest common multiple of the divisors in the above formula. Some care must be taken in downstream applications since the resulting counts are no longer true counts, but rescaled probabilities. The main consequence here is that systems which use +1 smoothing are reconfigured to use +1000 smoothing.

2.3 Word embeddings

Table 1 summarises the word embeddings used here. Due to the large number of word forms a Finnish lemma can take, it is of note here whether the word embedding represents word forms or lemmas. In the case an embedding represents word forms, it is additionally of note whether it uses any subword or character level information during its training, which should help to combat data sparsity. Despite the use of subword information, none of these embeddings can analyse out of vocabulary word forms. Cross-lingual word embeddings embed words from multiple languages in the same space, a property utilised in Section 3.2.2.

To extend word representations to sequences of words such as sentences, taking the arithmetic mean of word embeddings (AWE) has been commonly used as a baseline. Various incremental modifications have been suggested. Rücklé et al. (2018)

[5]Made available at `https://github.com/frankier/fiwn`.

Table 2: Results of experiments

Family	System	Variant	F_1
Baseline	Limits	Floor	13.1%
		Ceiling	99.9%
	Random sense	-	29.8%
	MFS	-	50.4%
Knowledge	UKB	No freq	51.8%
		No freq + Extract	52.2%
		Freq	54.5%
		Freq + Extract	54.9%
	Cross-lingual Lesk	No freq	32.6% – 48.2%[a]
		Freq	48.2% – 52.4%[a]
Supervised	SupWSD	No embeddings	72.9%
		Word2Vec_{-s}	73.6%
		Word2Vec	73.1%
		fastText_{-s}	73.3%
		fastText	73.4%
	AWE-NN	—	72.9% – 75.8%[b]

[a] See Table 3
[b] See Table 4

suggest concatenating the vectors formed by multiple power means, including the arithmetic mean. Variants CATP3 and CATP4 are used here. The former is the concatenation of the minimum, arithmetic mean, and the maximum, while the latter contains also the 3rd power mean. Arora et al. (2017) proposed Smooth Inverse Frequency (SIF), by taking a weighted average according to $\frac{a}{a+\mathrm{p}(w)}$, where a is a parameter and $\mathrm{p}(w)$ is the probability of the word. Arora et al. (2017) perform common component removal on the resulting vector. In the variant used here, (referred to as pre-SIF) a is set to the suggested value of 10^{-3} and common component removal is not performed, while $\mathrm{p}(w)$ is estimated based upon the word frequency data of Speer et al. (2018)[6].

3 Systems and Results

This evaluation is based on the all-words variant of the WSD task. In this task, the aim is to identify and disambiguate all words in some corpus. This is contrasted with the lexical sample approach, where a fixed set of words are chosen for evaluation. There are many systems and approaches which have been proposed for performing WSD. To select techniques for this evaluation, the following criteria were used:

- Prefer techniques which have been used in previous evaluations for English.
- Prefer techniques with existing open source code that can be adapted.
- Apart from this, include also simple schemes, especially if they represent an approach to WSD not covered otherwise.

[6] https://github.com/LuminosoInsight/wordfreq

The last criterion has led to the inclusion of multiple techniques based upon representation learning, where some representation of words or groups of words is learned in an unsupervised manner from a large corpus. To perform WSD based on these representations a relatively simple classifier, such as a nearest neighbour classifier, is then used. This approach to WSD additionally acts as a grounded extrinsic evaluation of the quality of the representations. The results of the evaluation are summarised in Table 2, with variants of the *Cross-lingual Lesk* and *AWE-NN* systems broken down in Tables 3 and 4. The rest of this section describes each of the systems in more detail.

3.1 Baseline

We can define limits for the performance of the WSD systems. The floor is defined by the proportion of unambiguous test instances. It is the F_1 score obtained by a system which makes correct guesses for unambiguous instances and incorrect guesses for every other instance. The ceiling is for systems based upon supervised learning, and is the proportion of test instances for which the true sense exists in the training data. It is the F_1 score obtained by a system which correctly associated every item in the test data with the true class seen in the training data, and makes an incorrect guess for every other instance.

The *random sense* baseline picks a random sense by picking the first sense according to a version of FinnWordNet without the frequency data from Section 2.2 i.e. the original sense order in FinnWordNet is assumed to be random. This also gives us a rough estimate of the average ambiguity of the gold standard, $\frac{1}{29.8\%} \approx 3$. The *MFS* baseline also picks the first sense, but uses the estimated frequencies from Section 2.2.

3.2 Knowledge based systems

Knowledge based WSD systems use only information in the LKB. In almost all dictionary style resources, this can include the text of the definitions themselves. In WordNet style resources, this can include also the graphical structure of the LKB.

3.2.1 UKB

UKB (Agirre et al., 2014) is a knowledge based system, representing the graph based approach to WSD. Since it works on the level of synsets, the main algorithm is essentially language independent, with the candidate extraction step being the main language dependent component. UKB can also make use of language specific word sense frequencies.

As noted in Agirre et al. (2018), depending on the particular configuration, it is easy to get a wide range of results using UKB. The configurations used here are based on the recommended configuration given by Agirre et al. (2018). For all configurations, the *ppr w2w* algorithm is used, which runs personalised page rank for each target word. One notable configuration difference here is that the contexts passed to UKB are fixed to a single sentence. This is the same input as is given to the other systems in this evaluation. Variations with and without access to word sense frequency information are given, (freq & no freq) with the latter assumed to be similar to the configuration given in Raganato et al. (2017).

By default, the lemmas and POS tags in the contexts given to UKB are from the sense tagged instances of EuroSense. Since some instances have been filtered from

Table 3: Results for variants of Lesk with cross-lingual word embeddings

Freq	Embedding	Agg	No expand		Expand	
			No filter	Filter	No filter	Filter
No	fastText	AWE	37.6%	34.9%	40.1%	40.0%
		CATP3	37.5%	35.5%	45.9%	46.9%
		CATP4	37.2%	35.2%	44.0%	45.2%
		pre-SIF	35.3%	34.5%	41.8%	40.1%
	Numberbatch	AWE	34.3%	*32.6%*	33.1%	34.3%
		CATP3	35.9%	35.6%	47.0%	47.7%
		CATP4	35.6%	35.4%	45.5%	46.2%
		pre-SIF	33.3%	33.3%	35.3%	36.0%
	Concatenated	AWE	36.7%	33.1%	37.1%	38.3%
		CATP3	36.3%	35.1%	47.6%	**48.2%**
		CATP4	36.3%	35.3%	45.9%	46.6%
		pre-SIF	33.8%	33.9%	40.0%	39.1%
Yes	fastText	AWE	49.4%	49.5%	50.1%	50.1%
		CATP3	49.3%	*48.2%*	49.2%	49.1%
		CATP4	49.3%	48.3%	49.5%	49.4%
		pre-SIF	52.2%	52.2%	**52.4%**	52.3%
	Numberbatch	AWE	49.7%	49.9%	50.5%	50.1%
		CATP3	49.3%	48.7%	48.8%	49.0%
		CATP4	49.5%	49.1%	49.0%	49.2%
		pre-SIF	52.0%	51.9%	51.9%	51.9%
	Concatenated	AWE	49.4%	49.6%	50.6%	50.3%
		CATP3	49.2%	48.5%	48.9%	49.1%
		CATP4	49.3%	48.9%	49.1%	49.3%
		pre-SIF	52.3%	52.0%	51.6%	51.7%

EuroSense so as to retain high precision, it may that UKB is hamstrung by an insufficient context size. To increase the information in the context without extending it beyond the sentence boundary, a high recall, low precision lemma extraction procedure based on OMorFi is performed. The procedure (referred to in Table 2 as *extract*) adds to the context all possible lemmas from each word form, including parts of compound words, and also extracts multiwords that are in FiWN.

3.2.2 Lesk with cross-lingual word embeddings

A variant of Lesk, referred to hereafter as Lesk with cross-lingual word embeddings (Cross-lingual Lesk) is included to represent the gloss based approach to WSD. The variant presented here is loosely based upon Basile et al. (2014). The technique is a derivative of simplified Lesk (Kilgarriff and Rosenzweig, 2000) in that words are disambiguated by comparing contexts and glosses. For each candidate definition, the word vectors of each word in the definition text are aggregated to obtain a definition vector. The word vectors of the words in the context of the word being disambiguated are also aggregated to obtain a context vector. Definitions are then ranked from best to

worst in descending order of cosine similarity between their definition vector and the context vector. Frequency data (freq) can be incorporated by multiplying the obtained cosine similarities by the smoothed probabilities of the synset given the lemma.

Since the words in the context are Finnish, but the words in the definitions are English, cross-lingual word vectors are required. The embeddings used are fastText, Numberbatch and the concatenation of both. Other variations are made by the choice of aggregation function, choosing whether or not to only include words which occur in FiWN, and whether glosses are expanded by adding also the glosses of related synsets. The gloss expansion procedure follows Banerjee and Pedersen (2002, Chapter 6). The results are summarised in Table 3.

3.3 Supervised systems

Supervised WSD systems are based on supervised machine learning. Most typically in WSD a separate classifier is learned for each individual lemma.

3.3.1 SupWSD

SupWSD (Papandrea et al., 2017) is a supervised WSD system following the traditional paradigm of combining hand engineered features with a linear classifier, in this case a support vector machine. SupWSD is largely a reimplementation of It Makes Sense (Zhong and Ng, 2010), and as such uses the same feature templates and its results should be largely comparable. It was chosen over It Makes Sense since it can handle larger corpora.

All variants include the POS tag and local colocation feature templates, and the default configuration includes also the set of words in the sentence. Variants incorporating the most successful configuration of Iacobacci et al. (2016), exponential decay averaging of word vectors with a window size of 10, are also included for each applicable word embedding from Section 2.3. For each configuration incorporating word vectors, variants without the set of words in the sentence are included, denoted e.g. Word2Vec$_{-s}$.

3.3.2 Nearest neighbour using word embeddings

Nearest neighbour using word embeddings has been used previously by Melamud et al. (2016) as a baseline. This system is very similar to the one outlined in Section 3.2.2. The main difference is that word senses are now represented by all memorised training instances, each themselves represented by the aggregation of word embeddings in their contexts. When a training instance is the nearest neighbour of a test instance, based on cosine distance, its tagged sense is applied to the test instance. This moves the technique from the realm of knowledge based WSD to supervised WSD. Since both tagged instances and the untagged context to be disambiguated are in Finnish, the constraint that word embeddings must be cross-lingual is removed. The results are summarised in Table 4.

4 Discussion & Conclusion

This paper has presented the first comparative WSD evaluation for Finnish. In the results presented here, several systems beat the MFS baseline. Of the knowledge based systems, both UKB and some variants of cross-lingual Lesk incorporating frequency

Table 4: Nearest neighbour using word embeddings

	AWE	CATP3	CATP4	pre-SIF
fastText	74.1%	74.1%	74.2%	74.1%
Numberbatch	74.5%	75.0%	74.9%	74.3%
Word2Vec	73.6%	*72.9%*	73.1%	73.8%
Concat 2[a]	75.1%	**75.8%**	75.5%	75.0%
Concat 3[b]	73.9%	73.2%	73.4%	74.5%

[a] Concatenation of fastText and Numberbatch

[b] Concatenation of fastText, Numberbatch and Word2Vec

information managed to clear the baseline. All the supervised systems tested beat it by a 20% margin. For techniques incorporating aggregates of word vectors, CATP3 reliably outperformed a simple arithmetic mean across a variety of configurations.

This evaluation may be limited by a number of issues. Multiple issues stem from the use of EuroSense. Due to the way it is automatically induced, it contains errors, making its use problematic, especially its use as a gold standard. First we model these errors as occurring in an essentially random manner. In this case a perfect WSD system would get a less than perfect score, and in fact the performance of all systems would be expected to decrease. It is worth noting that since inter-annotator agreement can be relatively low for word sense annotation, manual annotations can also be modelled as having this type of problem to some degree. Random errors in the training data would also cause the supervised systems to perform worse, however this does not effect the overall integrity of the evaluation. However, it is likely that EuroSense in fact contains systematic errors. One type of systematic error is an error of omission: EuroSense assigns senses to a subset of all possible candidate words, filtering out those which the Babelfy algorithm cannot assign sufficient confidence to, meaning that the gold standard may be missing words which are in some sense more difficult, artificially increasing the score of systems which would also have problems with these same words. Perhaps worse are systematic errors which bias certain lemmas within certain types of contexts to certain incorrect senses. In this case, supervised systems may seem to perform better, but only because they are essentially learning to replicate the systematic errors in EuroSense rather than because they are performing WSD more accurately.

Another factor which may cause this evaluation to present too optimistic a picture of the performance of supervised systems is that the evaluation corpus and training corpus are from the same domain, parliamentary proceedings, which could result in an inflated score in comparison to an evaluation corpus from another domain. Finally, since the corpus is derived from EuroParl, the original language of most text is likely not Finnish. Particular features of translated language, sometimes referred to as translationese may affect the applicability of the results to non translated Finnish[7].

Finally, the MFS baseline may have been handicapped in terms of its performance. On the one hand, the MFS baseline may be reasonably analagous with MFS baselines in WSD evaluations for other languages in that it is ultimately derived from frequency data which is out of domain. On the other hand, estimating the frequencies based on English frequency data is likely quite inaccurate when compared to a possible estimation based on a reasonably sized Finnish language tagged corpus.

[7]For an exploration of some features of translationese in EuroParl, see Koppel and Ordan (2011).

Further work could address the issues with the gold standard by creating a cross-domain manually annotated corpus, ideally based on a corpus of text originally in Finnish. A training corpus could also be created manually, but this would be a much larger task. This would however allow a better MFS baseline to be created. A less work intensive way of improving the situation with the MFS baseline would be to add one based on the supervised training data, and consider this as an extra MFS baseline, only for supervised methods.

The implementations of the techniques reimplemented for this evaluation and the scripts and configuration files for the adapted open source systems are publicly available under the Apache v2 license. To ease replicability further, the entire evaluation framework, including all the requirements, WSD systems and lexical resources are made available as a Docker image[8].

Acknowledgments

Thanks to the anonymous reviewers for their useful comments. Thanks also to my wife Miia for helping with the Finnish abstract. Finally, thanks to my supervisor Michael Cochez for his valuable advice and comments.

References

Eneko Agirre, Oier López de Lacalle, and Aitor Soroa. 2018. The risk of sub-optimal use of open source nlp software: Ukb is inadvertently state-of-the-art in knowledge-based wsd. *arXiv preprint arXiv:1805.04277* .

Eneko Agirre, Oier López de Lacalle, and Aitor Soroa. 2014. Random walks for knowledge-based word sense disambiguation. *Computational Linguistics* 40(1):57–84.

Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2017. A simple but tough-to-beat baseline for sentence embeddings .

Satanjeev Banerjee and T Pedersen. 2002. *Adapting the Lesk algorithm for word sense disambiguation to WordNet*. Master's thesis, University of Minnesota.

Pierpaolo Basile, Annalina Caputo, and Giovanni Semeraro. 2014. An enhanced lesk word sense disambiguation algorithm through a distributional semantic model. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. pages 1591–1600. http://www.aclweb.org/anthology/C14-1151.

Claudio Delli Bovi, Jose Camacho-Collados, Alessandro Raganato, and Roberto Navigli. 2017. Eurosense: Automatic harvesting of multilingual sense annotations from parallel text. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. volume 2, pages 594–600.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087* .

[8] `https://github.com/frankier/finn-wsd-eval`

Philip Edmonds. 2002. Senseval: The evaluation of word sense disambiguation systems. volume 7. http://www2.denizyuret.com/ref/edmonds/edmonds2002-elra.pdf.

Murhaf Fares, Andrey Kutuzov, Stephan Oepen, and Erik Velldal. 2017. Word vectors, reuse, and replicability: Towards a community repository of large-text resources. In *Proceedings of the 21st Nordic Conference on Computational Linguistics, NoDaLiDa, 22-24 May 2017, Gothenburg, Sweden*. Linköping University Electronic Press, Linköpings universitet, 131, pages 271–276.

Ignacio Iacobacci, Mohammad Taher Pilehvar, and Roberto Navigli. 2016. Embeddings for word sense disambiguation: An evaluation study. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. volume 1, pages 897–907.

Neeme Kahusk, Heili Orav, and Haldur Oim. 2001. Sensiting inflectionality: Estonian task for senseval-2. In *Proceedings of SENSEVAL-2 Second International Workshop on Evaluating Word Sense Disambiguation Systems*. Association for Computational Linguistics, pages 25–28. http://www.aclweb.org/anthology/S01-1006.

Adam Kilgarriff and Joseph Rosenzweig. 2000. English senseval: Report and results. In *LREC*. volume 6, page 2.

Moshe Koppel and Noam Ordan. 2011. Translationese and its dialects. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, pages 1318–1326.

Krister Lindén and Lauri Carlson. 2010. Finnwordnet–finnish wordnet by translation. *LexicoNordica–Nordic Journal of Lexicography* 17:119–140. http://www.ling.helsinki.fi/ klinden/pubs/FinnWordnetInLexicoNordica-en.pdf.

John C. Mallery. 1988. *Thinking About Foreign Policy: Finding an Appropriate Role for Artificially Intelligent Computers*. Master's thesis, Massachusetts Institute of Technology.

Oren Melamud, Jacob Goldberger, and Ido Dagan. 2016. context2vec: Learning generic context embedding with bidirectional lstm. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*. pages 51–61.

Márton Miháltz. 2010. *Semantic resources and their applications in Hungarian natural language processing*. Ph.D. thesis, Pázmány Péter Katolikus Egyetem.

George A Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J Miller. 1990. Introduction to wordnet: An on-line lexical database. *International journal of lexicography* 3(4):235–244.

George A Miller, Claudia Leacock, Randee Tengi, and Ross T Bunker. 1993. A semantic concordance. In *Proceedings of the workshop on Human Language Technology*. Association for Computational Linguistics, pages 303–308.

Andrea Moro, Alessandro Raganato, and Roberto Navigli. 2014. Entity Linking meets Word Sense Disambiguation: a Unified Approach. *Transactions of the Association for Computational Linguistics (TACL)* 2:231–244. http://www.aclweb.org/anthology/Q14-1019.

Roberto Navigli and Simone Paolo Ponzetto. 2012. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence* 193:217–250. https://doi.org/10.1016/j.artint.2012.07.001.

Simone Papandrea, Alessandro Raganato, and Claudio Delli Bovi. 2017. Supwsd: A flexible toolkit for supervised word sense disambiguation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. pages 103–108.

Tommi A Pirinen. 2015. Development and use of computational morphology of finnish in the open source and open science era: Notes on experiences with omorfi development. *SKY Journal of Linguistics* 28:381–393.

Alessandro Raganato, Jose Camacho-Collados, and Roberto Navigli. 2017. Word sense disambiguation: A unified evaluation framework and empirical comparison. In *Proceedings of EACL*. pages 99–110. https://aclanthology.info/pdf/E/E17/E17-1010.pdf.

Andreas Rücklé, Steffen Eger, Maxime Peyrard, and Iryna Gurevych. 2018. Concatenated p-mean word embeddings as universal cross-lingual sentence representations. *arXiv preprint arXiv:1803.01400* .

Miikka Silfverberg, Teemu Ruokolainen, Krister Lindén, and Mikko Kurimo. 2016. Finnpos: an open-source morphological tagging and lemmatization toolkit for finnish. *Language Resources and Evaluation* 50(4):863–878.

Robert Speer, Joshua Chin, and Catherine Havasi. 2016. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge .

Robyn Speer, Joshua Chin, Andrew Lin, Sara Jewett, and Lance Nathan. 2018. Luminosoinsight/wordfreq: v2.2. https://doi.org/10.5281/zenodo.1443582.

Alan M Turing. 1950. Computing machinery and intelligence. *Mind* 59(236):433–460. https://doi.org/10.1093/mind/LIX.236.433.

Kadri Vider and Kaarel Kaljurand. 2001. Automatic wsd: Does it make sense of estonian? In *Proceedings of SENSEVAL-2 Second International Workshop on Evaluating Word Sense Disambiguation Systems*. Association for Computational Linguistics, pages 159–162. http://www.aclweb.org/anthology/S01-1039.

David Yarowsky, Silviu Cucerzan, Radu Florian, Charles Schafer, and Richard Wicentowski. 2001. The johns hopkins senseval2 system descriptions. In *The Proceedings of the Second International Workshop on Evaluating Word Sense Disambiguation Systems*. Association for Computational Linguistics, pages 163–166.

Zhi Zhong and Hwee Tou Ng. 2010. It makes sense: A wide-coverage word sense disambiguation system for free text. In *Proceedings of the ACL 2010 system demonstrations*. Association for Computational Linguistics, pages 78–83.