Speech
Etymology
Idioms
Glossary
NLP
Lemma
Meaning
Corpora
Dictionary
Word
Lexicon
Definition
Headword
Pronunciation
Entry
Lexicology
Lexical Resources
Dictionary Use

λ EURALEX XIX

Congress of the European Association for Lexicography

Lexicography for inclusion

7-11 September 2021
Ramada Plaza Thraki
Alexandroupolis, Greece

www.euralex2020.gr

Proceedings Book
Volume 1

2020 Edition

Determining Differences of Granularity between Cross-Dictionary Linked Senses

Kouvara E., Gonzàlez M., Grosse J., Saurí R.

Oxford University Press, United Kingdom

Abstract

Linking dictionaries at the sense level is highly beneficial because it facilitates the mutual enhancement of the linked datasets or the possibility of deriving new products from the combination of the two. However, one of the greatest challenges in cross-dictionary sense linking is that linked senses, although referring to the same meaning, may actually differ in their semantic extent due to dictionary distinctions of sense granularity. Not every pair of linked senses is therefore qualitatively the same. However, being able to identify and classify these differences is a crucial step towards enabling the comprehensive exploitation of sense-linked datasets. In this paper, we present a system to automatically identify the relation of sense links between a bilingual and a monolingual dictionary. Using sense granularity annotations by lexicographers as the gold standard, we trained a machine learning model to classify the relation between cross-dictionary linked senses as one of the following categories: *perfect*, where each sense fully covers the other sense; *wider/narrower*, where one sense fully encloses the other but not vice versa; *partial*, where each sense partially covers the other sense. Cross-validation shows the machine learning model to yield an overall accuracy of 86%, with a macro precision of 83% and a macro recall of 65% across the different classes. The model significantly outperforms a rule-based algorithm serving as the baseline.

1 Introduction

This paper presents XD-SemOver (Cross-Dictionary Semantic Overlap Classifier), a machine learning-based tool for automatically predicting the type of the relationship between two senses from different dictionaries that have been linked as referring to the same meaning.

A sense link consists of a pair of senses, each from one of the linked dictionaries, belonging to the same lexeme and referring to the same meaning. Different dictionaries, however, may diverge in how they split the meaning of the words into different senses, that is, in the sense granularity criteria they apply. As a result, the two senses in a sense link may not be fully equivalent; rather, the following can hold:

- *Sense inclusion.* One sense has a wider (or narrower) semantic extent than the other. That is, one sense fully includes the other one.
- *Partial overlap.* The two senses overlap to some degree, but each of them includes a meaning component which is not covered by the other.
- *Perfect match.* The two senses refer exactly to the same meaning.

These distinctions are essential to ensure accuracy and completeness of information when inferring new multilingual data from chains of senses across languages. Otherwise, there is the risk of omitting important information that may be present in one of the senses, but not on the other. Assume, for example, the following chain of equivalent senses $S_{tA} - S_{tB} - S_{tC}$, as presented in the dictionaries for languages A, B and C. Each of these senses is associated with a term in its corresponding language: t_A, t_B, and t_C. If sense S_{tA} for term t_A in language A has a narrower semantic extent than sense S_{tB}, and S_{tB} is also semantically wider than S_{tC}, then it is not possible to safely infer that term t_C in language C is a translation of term t_A in language A.

This paper presents a component for automatically classifying differences in *sense granularity* between two senses from different dictionaries that refer to the same meaning, thereby establishing how the senses diverge with regards to their *semantic extension*. It has been developed as a component in a suite of tools for rich sense linking across dictionaries, which also includes: the sense linking system described in Saurí et al. (2019), a quality estimator tool that generates certainty values for automatically linked pairs (Grosse & Saurí 2020), and an annotation tool supporting the creation of manually labelled data for developing those resources (Gonzàlez, Buxton & Saurí 2020). On top of these tools, we are currently developing a system for automatically inferring new bilingual content from cross-dictionary sense links.[1] Thus, our purpose is to take previous work on sense linking one step further, therefore contributing to a higher quality for new bilingual content generation.

The paper is structured as follows. Section 2 reviews previous work that relates to the one presented here. Then, section 3 presents the overall project methodology, which is further detailed in sections 4 to 6. Results are discussed in section 7. Section 8 closes with final remarks and suggestions for future work.

---

[1] This project as a whole is part of a wider programme aiming to create multilingual data by combining Linguistic Linked Open Data and language technologies (Prêt-à-LLOD: https://www.pret-a-llod.eu/).

2   Related Work

The goal of this project is to determine the semantic overlap between two dictionary senses that refer to the same meaning, which directly touches on the notion of word polysemy and one of its central issues, namely, sense granularity. The critical question is how to split the different uses of a polysemous word into discrete senses. Within the lexical semantics field, much discussion has revolved around the notion of the meaning of a word and its analysis into different senses (e.g. Apresjan 1973; Cruse 1986; Pustejovsky 1995; Hanks & Pustejovsky 2005; Hanks 2013). From the more applied view of lexicography, the criteria for splitting senses as part of the dictionary creation process has also been an important topic of debate. Among many others, see for instance, Shcherba 1940/1995; Stock 1984; Wierzbicka 1985; Geeraerts 1990; Sinclair 1991; Fillmore & Atkins 1994; Kilgarriff 1997; Atkins & Rundell 2008.

At the computational arena, sense granularity has an impact in at least two areas of work: word sense disambiguation (WSD) on the one hand, and sense linking (or alignment) on the other. Within WSD, a task that in the past decades has been mostly tackled using machine learning-based approaches, sense granularity poses challenges when a word is split into excessively granular senses. The greater the granularity, the higher the complexity of the system's learning job. One solution comes from clustering senses to obtain more coarse-grained sense inventories (Navigli 2006; Navigli et al. 2007; Cinková, Holub & Kríž 2012). More recently, a data-driven approach grounded in distributional semantics (e.g. Erk 2010) and which is now benefiting from recent advances in deep learning (e.g. Pilehvar & Camacho-Collados 2019; Breit et al. 2020), has been moving away from defining senses a priori, thus circumventing the problem of sense granularity.

Within the computational framework, sense granularity also poses challenges to the activity around automatic sense linking (or alignment), which is precisely the area of work of our project. The overall goal of this area is to connect lexical databases (including dictionaries) at the sense level. Doing so enriches lexical content and enables the creation of new resources. There has been much research on this topic over the past two decades. Some work has been applied to the linking of lexical knowledge bases such as WordNet (Miller 1998) or Wikipedia,[2] which as opposed to traditional dictionary datasets, organise their content in a graph-based structure representing lexical relations among senses (Gurevych, Eckle-Kohler & Matuschek 2016). Other sense linking approaches have focused on word overlap, i.e. the number of common words among sense definitions (Ponzetto & Navigli 2010), while others have taken advantage of similarity distance measures (Ruiz-Casado, Alfonseca & Castells 2005; Ahmadi, Arcan & McCrae 2019). More recently, other approaches have explored using machine learning techniques to align senses between two dictionaries (Saurí et al. 2019).

Sense linking (or alignment) has also taken place across lexical resources of different languages in order to support cross-lingual information retrieval tasks (e.g., Gollins & Sanderson 2001; Massó et al. 2013) or to facilitate the rapid creation of new bilingual or multilingual lexicons, thus touching on the area of lexical translation (Varga, Yokoyama & Hashimoto 2009; Mausam et al. 2009; Wushouer et al. 2014; Villegas et al. 2016; Ordan et al. 2017; Gracia et al. 2019, among others).

Most of those approaches to sense linking, however, do not take into account the issue of sense granularity; namely, the possibility that the senses from two different sources that have been aligned as corresponding to the same meaning differ in their semantic extension (i.e. one has a wider or narrower semantic reference than the other). As argued in the introduction, these distinctions are critical because when generating new content from the alignment of two senses, there is a risk of leaving out crucial semantic information that is only present in one of the two senses. Nevertheless, to the best of our knowledge, there is only one sense linking project that has modelled these distinctions as part of the knowledge to be learnt: the ELEXIS Monolingual Word Sense Alignment Shared Task.[3] Aware of the relevance of this information, our sense linking project also includes a component for automatically identifying these distinctions. To date, however, there is no information published on the ELEXIS task for us to compare approaches or results.

3 Methodology

We developed XD-SemOver from a supervised machine learning approach following the standard methodology:

1. Delimiting the problem, i.e. determining the relevant distinctions that need to be learnt by the classifier.
2. Developing the dataset to be used for training and testing the system. For this, lexicographers annotated sense links using the categories determined in the previous step. The annotation was carried out using XD-AT, a Cross-Dictionary Annotation Tool supporting the manual categorisation of differences in sense granularity between two linked senses (Gonzàlez, Buxton & Saurí 2020).
3. Training and evaluating a classifier. We experimented with several models, namely a boosted decision trees algorithm, a multi-layer perceptron, and a pre-trained BERT-model. We also evaluated different settings (i.e. data balancing and parameter tuning) to identify the best approach.

The following sections detail the development carried out towards each of these aspects.

4 Delimiting the Problem

We distinguished the different types of sense links based on two kinds of relationships that may hold:

---

[2] https://www.wikipedia.org/
[3] https://competitions.codalab.org/competitions/22163#learn_the_details-overview

Overlapping: Determined by the extension of the meaning expressed by sense S_A that aligns with sense S_B. It can be:
- *Full*: S_A fully overlaps with S_B when S_A expresses the entire meaning of S_B.
- *Partial*: S_A partially overlaps with S_B when S_A only expresses part of the meaning of S_B.

Enclosing: Determined by whether S_A covers the entire extension of sense S_B. It can be:
- *True*: S_A encloses S_B if S_A covers the entire extension of S_B.
- *False*: S_A does not enclose S_B if S_A covers not all but only part of S_B.

These two levels of description can be orthogonally combined, generating a 4-fold distinction: *perfect*, *narrower-than, wider-than*, and *partial*. *Perfect* indicates that each sense aligns completely throughout the full extension of the other one. In other words, each sense fully covers the other. In contrast, *narrower-than* and *wider-than* account for sense pairs where a sense in one dictionary has a broader meaning than the sense in the other dictionary. It occurs when the meaning of one sense fully overlaps with the other one but does not fully enclose it (*narrower-than*), or the other way around (*wider-than*). Finally, *partial* denotes that each sense extends beyond the reference of the other. In this case, each sense includes a meaning that is not covered by the other. Table 1 illustrates the four relations that can occur between two linked senses. For the annotation task, we used these four sense link type classes.

The idea and the vocabulary for this distinction relate to the way the Simple Knowledge Organization System (SKOS) (Miles & Bechhofer 2009) expresses exact or fuzzy matching of concepts from one scheme to another using broader-narrower, or associative relationships. They also show strong similarities to the 5-fold categorisation in the ELEXIS Monolingual Word Sense Alignment Task (McCrae 2020), where the categories used are: *exact*, *broader*, *narrower*, *related*, or *none*. In our case, the *none* category is irrelevant, because we omitted from the classification process any sense pairs that are not linked.

	Perfect match	Different sense granularity		Different sense boundaries
Meaning alignment	S_A ——— S_B ———	S_A — S_B ———	S_A ——— S_B —	S_A ——— S_B ———
Grounding relationships				
S_A overlapping with S_B	fully	fully	partially	partially
S_A enclosing S_B	yes	no	yes	no
Sense link types	Perfect	Narrower-than	Wider-than	Partial
Symbol	=	<	>	~

Table 1: Types of sense links.

5 Gold Standard

5.1 Dataset Sampling

The dataset used to create the gold standard for developing the system includes the sense links between The Oxford Dictionary of English and three bilinguals published by Oxford University Press: English to Spanish, English to Russian, and English to Chinese (EN-ES, EN-RU, and EN-ZH). This set consists of 210,148 sense pairs that had been previously linked by human annotators.

Lexemes in this collection were split according to:
1) Their lexical category (noun, verb, adjective, adverb/preposition, or other). We could observe that each of these classes present a different behaviour in what refers to polysemy, and therefore pose different issues to sense link annotation.
2) Their polysemy degree (single-sense, small-size, medium-size, and large-size entry). Similar to above, senses in highly polysemous lexemes are more challenging to align than those in, e.g., monosemous entries.

With this classification, we aimed to help annotators focus on similar cases at a time, presumably with a similar degree of difficulty and similar features. Thus, for the manual annotation effort, we created batches of 100 sense links with the same lexical category and polysemy degree. Also, all sense links belonging to a lexeme were put together into the same batch.

Due to resource constraints, we could only annotate a subsample of the original dataset. For subsampling, we calculated the percentage of links for each combination of lexical category and polysemy degree and extracted 5% of the full collection of links. The sample sums up to 10,919 links (3,965 from EN-ES, 2,445 from EN-RU, and 4,403 from EN-ZH). Furthermore, we added additional batches that were annotated by several annotators (so-called *shared batches*) to compute inter-annotation agreement (IAA).

5.2 Manual Annotation Effort

Four annotators carried out the granularity classifications, and a fifth one acted as the judge to resolve disagreements in the *shared* batches, all of them expert lexicographers. Resolving disagreements was essential to obtain the labels conforming to the gold standard.

Sense links were classified according to the 4-fold distinction presented in section 4: *perfect match*, *wider-than*, *narrower-than*, and *partial match*. Additionally, annotators could use the tag *unlink*, if they considered that the sense pair did not correspond to a link, and *donotknow*, if they were uncertain about it.

The lexicographers annotated a total of 15,577 sense links, organised into 160 batches. Of these batches, 146 were annotated by one person (52, 40 and 54 batches from the EN-ES, EN-RU and EN-ZH dictionaries respectively). The remaining 14 batches were *shared* batches annotated by multiple annotators (5 EN-ES, 4 EN-RU and 5 EN-ZH), resulting in a total of 20,628 annotations.[4]

5.3 Inter-annotation Agreement

We used the annotations in the *shared* batches to analyse inter-annotator agreement as a proxy for the difficulty of the task. Since we are using the annotations as the gold standard for training the classifier, the inter-annotator agreement also represents the upper-bound for the classifier's performance. We measured the inter-annotator agreement using Fleiss' kappa metric (Fleiss 1971). The data consisted of the sense links annotated by all lexicographers involved, and included sense links labelled *donotknow* and *unlink*. Table 2 shows the results by dictionary, polysemy degree and lexical category.[5] Note that a kappa score of 0 signifies agreement that is fully explained by chance while a kappa score of 1 implies perfect agreement. According to Landis & Koch (1977), kappa scores between 0.41 and 0.60 denote moderate agreement and scores between 0.61 and 0.80 show substantial agreement.

			EN-ES			EN-RU			EN-ZH		
	Polysemy degree:		**S**	**M**	**L**	**S**	**M**	**L**	**S**	**M**	**L**
lexical category		adjective	0.62			0.55			0.64		
		adverb/preposition	0.72			0.64			0.66		
		noun		0.67			0.59			0.59	
		verb			0.48			0.50			0.42
		other	0.61						0.60		

Table 2: Fleiss' kappa inter-annotator agreement by lexical category, polysemy degree and dictionary.

Batches containing lexemes with lower polysemy degree show higher agreement, which suggests that these sense links were easier to classify. Likewise, the disagreement increases as the polysemy degree gets higher in all three languages. In addition, all three languages obtain similar kappa scores on average, but we observe differences for lexical categories. Adverb/prepositions get the highest values across the 3 datasets. Kappa values for adjectives drop significantly in EN-ES and EN-RU compared to adverb/preposition, while they are similar in EN-ZH. Verbs seem to carry less complexity in EN-RU: they only lose 9.28 points with respect to the average, compared to 18.15 loss in EN-ZH and 15.39 in EN-ES.

6 Building the Classifier

6.1 Models Explored

To develop our classifier, we experimented with three different approaches: using an AdaBoost ensemble of decision trees (Hastie et al. 2009), building a neural network (Glorot & Bengio 2010), and fine-tuning a pre-trained BERT model

---

[4] Before clean-up, the set comprised 14,178 unique annotations, 1,397 quadruple annotations (5,588 in total), and 862 judge reviews.
[5] Note that shared batches did not include links from monosemous lexemes, but only those in the polysemy degree classes of small (S), medium (M), and large (L).

(Devlin et al. 2018).[6] We also implemented a baseline model to be used as a benchmark in the evaluation of results. The following paragraphs outline each of these models.

6.1.1 AdaBoost

The first classification model we trained was an AdaBoost ensemble of decision trees. The main advantages of decision trees are that they are highly interpretable, they allow both numerical and categorical input data, and they are computationally efficient.

The implementation of AdaBoost selected for our task makes use of the meta-estimator AdaBoostClassifier,[7] provided by *sklearn*. The AdaBoost classifier uses an ensemble of DecisionTreeClassifiers[8] as base estimators. We experimented with a variety of parameters using grid search (GridSearchCV[9]), as described in section 6.2.

6.1.2 Neural Network

The second classification model we trained was a feedforward neural network. Neural networks are known to perform well with high dimensionality data, and they can model complex, non-linear relations between input variables.

For the implementation, we chose the MLPClassifier from *sklearn*.[10] Again, we employed grid search to fine-tune the hyperparameters, as described in section 6.2. Due to the MLPClassifier's sensitivity to feature scaling, we further experimented with feature normalisation and standardisation.

6.1.3 Deep Learning with BERT

The third model was developed by fine-tuning a pre-trained BERT model for the sense granularity classification task. BERT generates different word embeddings for the same word depending on its context, and words in the same context generate similar word embeddings. For this reason, BERT is a viable candidate for predicting the granularity of sense links.

An essential part of BERT is Next Sentence Prediction (NSP), where the model learns to understand sentence relations by learning whether or not a given sentence follows another (Devlin et al. 2018). Accordingly, we treat the sense link granularity task as an NSP task. By representing each sense as a sentence made up of its lexical information (e.g. its definition), we train the model to classify sense links based on the chained sentences representing the two senses.

6.1.4 Baseline

Finally, we also created a baseline classification method against which to compare the above models. The baseline model is a rule-based algorithm employing the method described in Table 3, which takes into account the number of times each sense is linked to a sense in the other dictionary.

Situation	Graphical representation	Resulting link type
Neither S_A nor S_B are linked to any other sense in the other dictionary	S_A ●——● S_B	**perfect (=)**
In addition to S_{B_1}, S_A is linked to one or more other senses (in blue) in the bilingual dictionary	S_A $S_{B\ 1}$ $S_{B\ j}$ $S_{B\ n}$	**wider-than (>)** (i.e. S_A is wider than S_{B_1})
In addition to S_{A_1}, S_B is linked to one or more other senses (blue) in the monolingual dictionary	$S_{A\ 1}$ $S_{A\ i}$ $S_{A\ n}$ S_B	**narrower-than (<)** (i.e. S_{A_1} is narrower than S_B)
Both S_{A_i} and S_{B_j} are linked to multiple senses in the other dictionary	$S_{A\ 1}$ $S_{B\ 1}$ $S_{A\ i}$ $S_{B\ j}$ $S_{A\ n}$ $S_{B\ n}$	**partial (~)**

Table 3: Baseline classification heuristics.

The algorithm works as follows: For senses S_A and S_B, it assigns the label: i) *perfect* if the number of links for both senses is equal to 1, ii) *wider-than* if the number of links is larger than 1 in S_A and equal to 1 in S_B, iii) *narrower-than* if the number of links is equal to 1 in S_A and larger than 1 in S_B, and iv) *partial* if the number of links for both senses is larger than 1.

---

[6] We used the Base, Uncased model, which can be downloaded from https://github.com/google-research/bert.
[7] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.AdaBoostClassifier.html
[8] https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html
[9] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html
[10] https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html

The rationale behind this method is that when a sense in one dictionary is linked to several senses in the other, it should have a wider semantic extension than any of those senses in the other dictionary. Thus, this one-to many relation is labelled as *wider-than*. Inversely, many-to-one cases are labelled as *narrower-than*. One-to-one relations are instances where both senses should refer to the same extent of meaning, thus being labelled *perfect*. Finally, many-to-many relations suggest that both senses include a meaning component not present in the other, therefore receiving the label *partial*.

6.2 Evaluation Method

To find the best hyperparameters for the two machine learning models, AdaBoost and NN, we used sklearn's GridSearchCV. The choice of values to test, shown in Tables 4 and 5, results from research on the algorithms and includes the default plus neighbouring orders of magnitude. The experimental results were evaluated using 10-fold cross-validation and compared against the baseline model's performance.

AdaBoost parameters	values
n_estimators	10, 100, 1000
learning_rate	1, 0.1, 0.01, 0.001
base_estimator__criterion	gini, entropy
base_estimator__max_depth	1, 10, None
random_state	999

Table 4: Hyperparameters grid for AdaBoost and Decision tree estimators.

Neural Network MLP parameters	values
hidden_layer_sizes	(100,), (72, 36), (72, 36, 18)
activation	tanh, relu
solver	adam
alpha	0.001, 0.0001, 0.00001
batch_size	10, 100, 1000
learning_rate_init	0.01, 0.001, 0.0001
early_stopping	True
random_state	999

Table 5: Hyperparameters grid for Neural Network (MLP).

6.3 Experiments

The experiments were organised in four rounds of iterations with different goals: The first three incrementally built on top of each other and focused on the machine learning algorithms (AdaBoost and NN), while the fourth involved fine-tuning a pre-trained BERT model. The following is a summary of the main characteristics of each round.

6.3.1 Round 1: Ground Base

In our initial experiment, we used the same 42 dictionary-based features that had been employed for training XD-BaSeLink, the sense linking classifier, to gain initial insights into the feasibility of the classification task. The features used for this round appear listed in the appendix of Saurí et al. (2019).

6.3.2 Round 2: Optimising Features and Hyperparameters

In this iteration, we introduced 56 additional features. These features were both binary and categorical, and they relate to the sense order, domain, register, region, definition/indicators, and examples of the senses in each pair. Furthermore, we added as a feature the number of links for each sense in the sense pair. This feature is used in creating the baseline model (see section 6.1.4), and it is intuitively informative for the classification task. Consider, for example, a sense pair (S_{mono}, S_{bil}) between a sense in the monolingual dictionary and another in the bilingual dictionary. Knowing that S_{mono} has

been linked to three senses in the bilingual dictionary and that S_{bil} has been linked twice intuitively makes it less likely that the sense link is *perfect*. Instead, and in agreement with the baseline algorithm, we may expect the link to fall into the *partial* category.

6.3.3 Round 3: Dataset Balancing

The dataset was highly imbalanced across the four classes, which can introduce a bias towards the majority classes. To counteract this bias, we experimented with several sampling techniques that balance out the classes.

Oversampling. The advantage of oversampling is that no observations (i.e. sense links) are lost. We used the Synthetic Minority Oversampling Technique for Nominal and Continuous (SMOTE-NC), which is a variation of SMOTE (Chawla et al. 2002) tailored for datasets with continuous and categorical features. By synthetically generating more instances of the minority class, inductive learners, like decision trees, can broaden their decision regions for that class. We assessed two variations of oversampling: Oversampling the minority classes to the full number of data points in the majority class and oversampling the minority classes to half of the number of data points in the majority class.

Undersampling. Similarly, we also experimented with undersampling the dataset, reducing the size of the majority class. For this, we used the Random Under Sampling (RUS) technique, which randomly selects and removes data points from the majority classes, while leaving the minority classes' samples intact, until a balanced distribution is achieved.

Hybrid resampling and boosting. The results for the previous iterations did not show a significant change in performance for the AdaBoost classifier, and thus we decided to experiment with a further classification strategy. We used two different implementations that combine the SMOTE oversampling and the random undersampling with the boosting techniques, instead of resampling the dataset first and then using the AdaBoost classifier. The first algorithm used, called SMOTEBoost (Chawla et al. 2003), is a hybrid sampling/boosting algorithm that creates synthetic examples from the minority class, thus indirectly changing the updating weights and compensating for skewed distributions. The second classifier, called RUSBoost (Seiffert et al. 2009), is a combination of a boosting algorithm that uses RUS to randomly remove examples from the majority class, giving a clear advantage in training time, in comparison to SMOTEBoost.

6.3.4 Round 4: BERT Experiments

We fine-tuned the pre-trained BERT model to the classification task. This approach requires text as input rather than features; hence we defined 7 different combinations of text input to represent both senses in the link. We selected that approach after the promising results obtained by Breit et al. (2020) in identifying the sense of a word as used in context. In their experiments, they try to determine whether the meaning of a word used in a particular context matches the target sense represented by either its definition or its hypernyms. This setup is related to ours since we also represented senses using their definition. We put words in context using sense examples, and we further characterise senses with their collocates, domain labels, and more. Also, we consider BERT as a proper representation model since it is a bidirectional neural network that learns to distinguish the meaning of a word depending on its context, hence it represents words appearing in similar contexts through similar word embeddings vectors.

7 Results and Discussion

We evaluated the classifiers using the following evaluation metrics: *accuracy*, *precision*, *recall*, and *f1-score*. We further differentiated these scores between *macro* (equal weight per class, higher relevance to the results for small classes), and *weighted* (weight-adjusted by class, greater relevance to larger classes).[11] *Macro* and *weighted* averages respectively provide lower and upper bounds to the true average of these metrics. We also looked at Cohen's kappa score (Cohen 1960) to measure agreement between the human annotation and the predicted labels. The best performing model consists of an AdaBoost ensemble of decision trees, which clearly outperformed both the neural network and the fine-tuned BERT model.[12] We trained it with the features used in Saurí et al. (2019) and additional features encoding the lexical category of the sense link lexeme and the number of links for each sense in the sense link. Table 6 and Table 7 show the results of this model compared to the baseline model presented in section 6.1.4.

The AdaBoost classifier outperforms the baseline model in global terms, with a macro averaged f1 score of 71% compared to the baseline of 65%. It also yields a higher overall accuracy of 86% compared to 80% for the baseline. Unsurprisingly, the majority class *perfect* consistently scores highest, whereas the minority class *partial* scores lowest. This tendency is also represented by the weighted metrics, which receive higher scores than the macro averages, indicating an overrepresentation of the majority class in prediction.

---

[11] We used weighted instead of micro because the task is a multiclass classification problem, for which the micro average would yield the same values for precision, recall and subsequently f1. Refer to: https://simonhessner.de/why-are-precision-recall-and-f1-score-equal-when-using-micro-averaging-in-a-multi-class-problem/ (05/2020)

[12] The best model, chosen based on weighted f1-score, was obtained with the following parameters: base_estimator_criterion=entropy, base_estimator_max_depth=1, learning_rate=0.05, n_estimators=100.

	Precision		recall		f1 score		class
	baseline	AdaBoost	Baseline	AdaBoost	baseline	AdaBoost	size
Performance by class							
narrower-than	0.81	**0.94**	0.65	0.65	0.73	**0.77**	2979
partial	0.31	**0.70**	**0.43**	0.27	0.36	**0.39**	494
perfect	**0.88**	0.85	0.89	**0.99**	0.89	**0.92**	8826
wider-than	0.58	**0.84**	**0.72**	0.70	0.64	**0.76**	1416
Overall performance							
macro avg	0.65	**0.83**	**0.67**	0.65	0.65	**0.71**	N/A
weighted avg	0.82	**0.87**	0.80	**0.86**	0.81	**0.85**	N/A

Table 6: Overall and per class performance scores for baseline and AdaBoost models. The last column (GS class size) provides an indication of the biased nature of the gold standard (GS)

	Baseline	AdaBoost classifier
accuracy	0.80	**0.86**
kappa	0.63	**0.71**

Table 7: Overall accuracy and kappa scores for baseline and AdaBoost models.

The confusion matrix in Figure 1 visualises the data bias: most incorrect classifications had the label *wider-than* or *narrower-than*, but were predicted *perfect*. The bias stems from the skewed training data, with the majority class *perfect* having almost 18 times as many samples as the minority class *partial*. The classes *wider-than* and *narrower-than* perform a bit better than *partial*, but worse than *perfect*. We attempted to address the present bias by synthetically resampling the training data. However, doing so decreased the overall performance unjustifiably in terms of accuracy and macro averaged f-1 score, which is why we chose to stick to the original dataset.
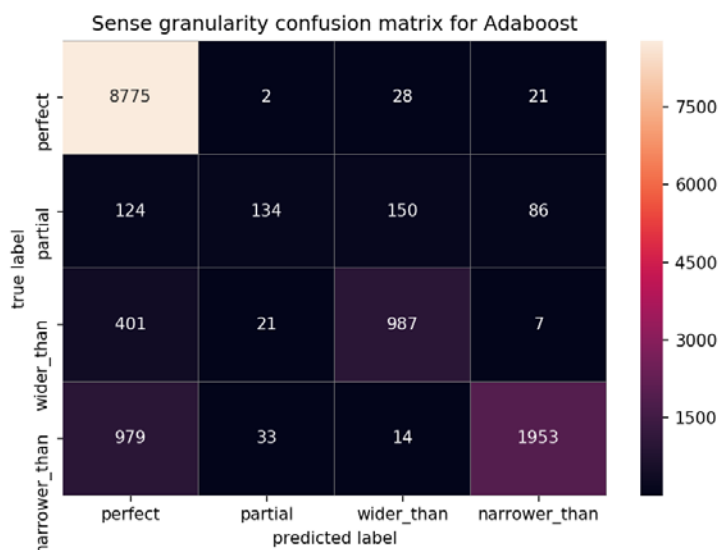

Figure 1: Heatmap showing the confusion matrix of predicted vs correct labels for the AdaBoost classifier.

As mentioned above, the training data is heavily imbalanced, posing the question whether the overall performance of our classifier could be improved by introducing additional (non-synthetic) data points, especially for the less represented classes. Alternatively, the algorithm may benefit from the introduction of additional features, which may aid in the identification of data points from the minority classes. Both of these approaches may prove viable alternatives to the sampling methods explored in this paper to decrease the algorithm's bias towards the majority class.

Overall, these results can be assessed as positive. We consider the *macro precision*, *weighted precision* and *weighted recall* values as quite good given that they are above 0.85 and reaching towards a 0.9, whereas the *macro recall* results are lower but still within a decent range. The lower *macro recall* was, however, expected, since the *macro average* metrics are very sensitive to class imbalance and assign greater prevalence to the smallest classes, for which we observed particularly low scores.

8 Conclusion

In this paper, we motivated the need for identifying distinctions of sense granularity between linked senses from different dictionaries to support automatic tasks in different areas, most significantly the creation and enhancement of multilingual lexical resources. As a solution, we proposed an automatic classifier that uses a standard supervised machine learning approach for multiclass classification. The resulting model performs convincingly, and can, therefore, be used for tagging linked senses with a reasonable degree of confidence. A limitation of the model is that it shows a bias towards predicting

the majority classes, due to the imbalanced nature of the training dataset. Future work can address this shortcoming in two ways: Firstly, by obtaining more manual annotations, especially for the minority classes, to have a balanced training dataset and thus improve the classifier's performance. And secondly, by enriching the training dataset with additional features that can help the classifier to distinguish among classes more accurately.

9 References

Ahmadi, S., M. Arcan, J. McCrae (2019). Lexical Sense Alignment using Weighted Bipartite b-Matching. In *Proceedings of the Poster Track of LDK 2019*, pp. 12-16.

Apresjan, J. D. (1974). Regular polysemy. *Linguistics*, *12*(142), pp. 5-32.

Atkins, B. T. S., M. Rundell (2008) *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.

Breit, A., Revenko, A., Rezaee, K., Pilehvar, M.T., Camacho-Collados, J. (2020). WiC-TSV: An Evaluation Benchmark for Target Sense Verification of Words in Context. URL: https://arxiv.org/abs/2004.15016v1

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, pp. 321-357.

Chawla, N. V., Lazarevic, A., Hall, L. O., & Bowyer, K. W. (2003). SMOTEBoost: Improving prediction of the minority class in boosting. In *European conference on principles of data mining and knowledge discovery* (pp. 107-119). Springer, Berlin, Heidelberg.

Cinková, S., Holub, M., & Kríž, V. (2012). Optimizing semantic granularity for NLP-report on a lexicographic experiment. In *Proceedings of the 15th EURALEX International Congress*, pp. 523-531.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psych. measurement*, *20*(1), pp. 37-46.

Cruse, D. A. (1986). *Lexical Semantics*. Cambridge: Cambridge University Press.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding.

Donandt, K., Chiarcos, C. (2019). Translation inference through multi-lingual word embedding similarity. In *Proc. of TIAD-2019 Shared Task Translation Inference Across Dictionaries, at 2nd Language Data and Knowledge (LDK) conference. CEUR-WS (May 2019)*.

Fillmore, C. J., Atkins B. T. S. (1994). Starting Where the Dictionaries Stop: The Challenge for Computational Lexicography. In B. T. S. Atkins and A. Zampolli (eds.) *Computational Approaches to the Lexicon*. New York: Oxford University Press, pp. 349–393.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*, 378-382.

Geeraerts, D. (1990). The Lexicographical Treatment of Prototypical Polysemy. In S. L. Tsohatzidis (ed.) *Meanings and Prototypes*. London: Routledge, pp. 195–210.

Glorot, X., Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 249-256.

Gollins, T., M. Sanderson (2001). Improving Cross Language Retrieval with Triangulated Translation. In *24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '01, pp. 90–95.

Gonzàlez, M., Buxton, C., Saurí, R. (2020). XD-AT: A Cross-Dictionary sense Alignment and mark-up Tool. In *Proceedings of the XIX EURALEX conference*. Alexandroupolis, Greece. To appear.

Grosse, J., Saurí, R. (2020). Principled Quality Estimation for Dictionary Sense Linking. In *Proceedings of the XIX EURALEX conference*. Alexandroupolis, Greece. To appear.

Gracia, J., B. Kabashi, I. Kernerman (2019). *Proceedings of TIAD-2019 Shared Task – Translation Inference Across Dictionaries*, co-located with the Language, Data and Knowledge Conference (LDK). Leipzig, Germany, May 2019.

Gurevych, I., J. Eckle-Kohler, & M. Matuschek (2016). *Linked Lexical Knowledge Bases: Foundations and Applications*. Morgan & Claypool Publishers.

Hanks, P. (2013). *Lexical Analysis: Norms and Exploitations*. MIT Press.

Hanks, P., Pustejovsky, J. (2005). A pattern dictionary for natural language processing. *Revue Française de linguistique appliquée*, *10*(2), pp. 63-82.

Hastie, T., Rosset, S., Zhu, J., & Zou, H. (2009). Multi-class adaboost. *Statistics and its Interface*, *2*(3), pp. 349-360.

Kilgarriff, A. (1997). I don't believe in word senses. *Computers and the Humanities*, *31*(2), pp. 91-113.

Landis, J. R., Koch, G. G. (1977). The measurement of observer agreement for categorical data. Biometrics, 33(1), pp. 159–174.

Massó, G., P. Lambert, C. Rodríguez-Penagos, & R. Saurí (2013). Generating New LIWC Dictionaries by Triangulation. In R. E. Banchs, F. Silvestri, T. Liu, M. Zhang, S. Gao, J. Lang, (eds.) *Information Retrieval Technology*, pp. 263–271.

Mausam, S. Soderland, O. Etzioni, D. Weld, M. Skinner, & J. Bilmes (2009). Compiling a Massive, Multilingual Dictionary via Probabilistic Inference. In *Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pp. 262–270.

McCrae, J. (2020, May). *ELEXIS Monolingual Word Sense Alignment Task*. Retrieved from CodaLab Competition: https://competitions.codalab.org/competitions/22163

Miles, A., Bechhofer, S. (2009). *SKOS Simple Knowledge Organization System Reference*. (W3C Recommendation). http://www.w3.org/TR/skos-reference/: World Wide Web Consortium.

Miller, G. A. (1998). *WordNet: An electronic lexical database*. MIT press.

Navigli, R. (2006). Meaningful clustering of senses helps boost WSD performance. In: *Proceedings of the 21st International Conference of Computational Linguistics and the 44th Meeting for the ACL*, pp. 105-112.

Navigli, R., K.C., Litkowski, O. Hargraves (2007). Semeval-2007 Task 07: Coarse-grained English all-words task. In *Proceedings of the 4th International Workshop on Semantic Evaluations*: 30-35. ACL.

Ordan, N., J. Gracia, M. Alper, I. Kernerman (2017). *Proceedings of TIAD-2017 Shared Task – Translation Inference Across Dictionaries. Language*, Data and Knowledge Conference, LDK 2017. Galway, Ireland, June 2017.

Pilehvar, M. T., Camacho-Collados, J. (2019). WiC: The word-in-context dataset for evaluating context-sensitive meaning representations. In *Proceedings of the NAACL-HLT Conference*. 2019.

Ponzetto, S. P., Navigli, R. (2010). Knowledge-rich word sense disambiguation rivaling supervised systems. In *Proceedings of the 48th annual meeting of the association for computational linguistics* (pp. 1522-1531). ACL.

Pustejovsky, J. (1995). *The Generative Lexicon*. Cambridge, Massachusetts: The MIT Press.

Ruiz-Casado, M., Alfonseca, E., & Castells, P. (2005). Automatic assignment of wikipedia encyclopedic entries to wordnet synsets. In *International Atlantic Web Intelligence Conference* (pp. 380-386). Springer, Berlin, Heidelberg.

Saurí, R., Mahon, L., Russo, I., & Bitinis, M. (2019). Cross-Dictionary Linking at Sense Level with a Double-Layer Classifier. In *2nd Conference on Language, Data and Knowledge (LDK 2019)*. Leibniz-Zentrum fur Informatik.

Shcherba, L. V. (1940/1995). Towards a general theory of lexicography. *Inter. Journal of Lexicography,* 8.4, pp. 314–350

Seiffert, C., Khoshgoftaar, T. M., Van Hulse, J., & Napolitano, A. (2009). RUSBoost: A hybrid approach to alleviating class imbalance. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, *40*(1), pp.185-197.

Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Stock, P. (1984). Polysemy. In R.K.K. Hartmann (ed.) *LEXeter '83 Proceedings*. Tübingen: Niemeyer, pp. 131–140.

Varga, I., S. Yokoyama, C. Hashimoto (2009). Dictionary generation for less-frequent language pairs using WordNet. In *Literary and Linguistic Computing*, Vol. 24, Issue 4, December 2009:449–466, https://doi.org/10.1093/llc/fqp025

Villegas, M., M. Melero, N. Bel, & J. Gracia (2016). Leveraging RDF Graphs for Crossing Multiple Bilingual Dictionaries. In *Proceedings of the Language Resources and Evaluation Conference, LREC 2016*, pp. 23–28.

Wierzbicka, A. (1985). *Lexicography and Conceptual Analysis*. Ann Arbor: Karoma.

Wushouer, M., D. Lin, T. Ishida, & K. Hirayama (2014). Pivot-Based Bilingual Dictionary Extraction from Multiple Dictionary Resources. In *PRICAI 2014: Trends in Artificial Intelligence*. Springer International Publishing, pp. 221–234.