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

Edited by Zoe Gavriilidou, Maria Mitsiaki, Asimakis Fliatouras

EURALEX Proceedings

ISSN 2521-7100

ISBN 978-618-85138-1-5

Edited by: Zoe Gavriilidou, Maria Mitsiaki, Asimakis Fliatouras

English Language Proofreading: Lydia Mitits and Spyridon Kiosses

Technical Editor: Kyriakos Zagliveris

2020 Edition

Principled Quality Estimation for Dictionary Sense Linking

Grosse J., Saurí R.

Oxford University Press, United Kingdom

Abstract

Estimating the quality of lexical data automatically linked on the sense level is challenging, as the quality of the predicted sense links can differ significantly across various datasets. This variability is especially problematic when quality estimation is limited to general statements about an extensive collection of sense pairs, such as the links between two entire dictionaries. We argue that estimating probabilities for individual sense pairs is a superior method for quality estimation for two reasons: Firstly, it allows us to draw more nuanced conclusions about the quality of linked lexical data. Secondly, it opens the door for merging automated with manual means of sense linking by pointing lexicographers towards sense pairs that are especially difficult to classify. We propose a method for generating such probability estimates for a supervised machine learning approach. We show that these probabilities successfully dissect the sense pairs based on the certainty of the classification algorithm, thereby enabling lexicographers to analyse and improve the quality of automatically linked lexical data effectively.

Keywords: Word sense linking; language data construction; semi-automated annotation; data quality estimation; probability estimation

1 Introduction

Research into automating lexicographic processes for the creation of dictionary content has gained significant traction in recent years. This research has been partly facilitated by the European Union, which currently funds two large programmes in the area: ELEXIS (European Lexicographic Infrastructure Programme)[1] and Prêt-à-LLOD.[2] One of the goals of the programmes is to link linguistic linked data on the sense level[3] and make such data a useful tool for real-life problem-solving. To date, automatically sense-linked datasets, dictionaries in specific, have been a topic of research but are rarely seen in the production context. The obstacles preventing the full leveraging of automated sense linking boil down to the issue of quality: automatically linked dictionaries do not yet satisfy the high-quality expectations of the market. At the same time, linking dictionaries manually, while ensuring superior link quality, takes a long time and is often unfeasible financially.

One potential solution to this problem is to combine automated means with human expertise to create cost-efficient high-quality linked datasets. A necessary step towards such semi-automated linking, we argue, is to quantify the quality of individual linked sense pairs in terms of probability estimates, as such probability estimates allow us to distinguish certain sense links from uncertain ones. Accordingly, in this paper, we describe an algorithm estimating the link probability of previously predicted sense links. These probability estimates allow us to more accurately predict the quality of newly generated sense links between two dictionaries. Additionally, they enable the joint application of automated and manual means of sense linking.

2 Related Work

There exist various approaches to linking lexical content on the sense level, including using a pivot language (e.g. Kaji, Tamamura, & Erdenebat 2008; Tanaka & Umemura 1994; Varga & Yokoyama 2009; Wushouer et al. 2014), translation graphs with circles (e.g. Alper, 2017; Mausam et al. 2009; Villegas et al. 2016) or weights (e.g. Proisl et al. 2017), dictionary triangulation (e.g. Gollins & Sanderson 2001; Massó et al. 2013), neural machine translation (e.g. Arcan et al. 2019), and supervised machine learning (e.g. Donandt, Chiarcos, & Ionov 2017; Saurí et al. 2019). Notably, even though there has been groundbreaking recent progress in machine translation and natural language processing, the 1994 seminal paper by Tanaka and Umemura remains a highly competitive baseline for dictionary linking on the sense level (Gracia, Kabashi, & Kernerman 2019). This observation, paired with the fact that researchers have explored numerous approaches, is a testament to the difficulty and uniqueness of the problem of sense linking. Therefore, we must ask: How viable is automated sense linking? What is the level of quality we can expect for a given set of sense links?

The diversity of approaches to sense linking necessitates a comparable diversity of approaches in quality estimation. While some have settled with extrapolating quality audits of modest data samples, others have aimed to rank potential sense links by obtaining a proxy for the certainty of the respective linking algorithm's prediction: Villegas et al. (2016) assume that the density of translation cycles corresponds to the likelihood of a correct sense link; Mausam et al. (2009) probabilistically evaluate possible paths in a translation graph to estimate the percentage of complete cycles, which in turn serves as the certainty proxy; Arcan et al. (2019) simply take the output of the machine translation algorithm as the

[1] https://elex.is/
[2] https://www.pret-a-llod.eu/
[3] Note that linguistic linked data often refers specifically to data in the RDF format. Here, however, we mean a format agnostic interpretation of linguistic linked data as language data that is linked on the sense level.

confidence score; Massó et al. (2013) quantify the reliability of a sense link by counting the number of dictionaries included in possible triangulations; Shezaf & Rappoport (2010) obtain certainty in the form of a similarity score for a pair of words in source and target language.

The previous work on linking lexical content described above, despite vast variations in approach, faces a common challenge: differentiating between different degrees of certainty for candidate sense links. As shown above, all approaches aiming to address the challenge make use of a proxy for the certainty that a given sense pair is a sense link.

The issue with the general method of estimating quality through proxies for certainty is that these proxies alone fall short of providing principled probability estimates. While the proxies are correlated with the underlying probabilities (after all this correlation is what makes the proxies valid in the first place), no previous work that we are aware of has convincingly quantified this correlation to estimate the underlying probabilities. Failing to obtain probability estimates means that, while (limited) general conclusions can be drawn about the quality of sufficiently large datasets, the actual probability of any individual predicted sense link being correct remains opaque. Furthermore, any conclusions drawn about the quality of a dataset based on a certainty threshold rely on the assumption that different datasets have an identical, or at least highly similar, distribution of certainty scores. However, due to the linguistic idiosyncrasies of languages, this assumption rarely holds. For instance, the degree of polysemy of words is (negatively) correlated with the certainty of prediction. This correlation means that we can expect languages with higher polysemy to result in less certain sense links than languages with lower polysemy. The lower certainty consequently results in lower data link quality, even if the certainty cut-off remains the same, rendering the original quality estimates imprecise.

3 Motivation

Supervised machine learning offers a promising opportunity for obtaining the desired probability estimates that solve the challenges mentioned above. For one, many machine learning algorithms, such as logistic regression, inherently output predictions in form of fractions in the unit interval. These outputs have widely been interpreted as probabilities. Even for machine learning algorithms where such an interpretation of the output is impossible or unprincipled, numerous calibration methods have been proposed for obtaining the underlying probabilities (e.g. Niculescu-Mizil & Caruana 2005). Therefore, machine learning approaches to sense linking lend themselves well to a natural quantification of link certainty, consequently allowing a ranking of sense pairs by link likelihood.

In our previous work, applying supervised machine learning to sense linking (Saurí et al. 2019), however, we are denied the straightforward ways of obtaining probability estimates. The reason lies within the dependency of the data. Defining the problem of sense linking as a binary classification task, thus making a machine learning approach feasible, entails assuming that each potential sense pair is independent of any other potential sense pair. This assumption is flawed; the link likelihood of any given sense pair depends, at least in part, on the connection the two senses have to other senses. Knowing that the two senses of a sense pair are already linked to other senses makes the sense pair less likely to be a link. Inversely, knowing that the two senses of a sense pair are not linked to any other senses makes the sense pair more likely to be a link. To account for the existing dependency, Saurí et al. (2019) employed a second algorithmic layer where the initial prediction of the machine learning algorithm is compared to a threshold quantifying the effect of the dependency across sense pairs. This additional classification layer significantly improves results. Unfortunately, it has the side effect that we can no longer take advantage of the probability estimates (pure or calibrated) that could be calculated in the first algorithmic layer because these estimates are tied to the assumption of independence between sense pairs. In other words, because the second algorithmic layer changes the predictions (therein accounting for the dependencies across sense pairs and improving results), the initial probability estimates lose validity. For instance, we find that a sense pair with an initial link probability of 0.3 may be classified a link because neither sense is linked to any other sense, while a sense pair with an initial link probability of 0.8 may be classified a non-link because both senses are already linked to numerous other senses.

This paper addresses these challenges by introducing an alternative method for obtaining principled link probabilities for Saurí et al. (2019).

4 Method

4.1 Starting Point

The algorithm for calculating probability estimates takes as input the results of the sense linker developed in Saurí et al. (2019). The results of the sense linker come in the form of a binary value for each sense pair: link or non-link. The sense linker predicts the binary category through a two-step process:

a. The first level of the algorithm applies a machine learning based classifier, returning an independent prediction *p* for each sense pair.
b. The second level of the algorithm determines the final classification (link or non-link) by comparing the prediction *p* from the first level to a threshold *t*. The threshold *t* is calculated for each sense pair using:
    - The number of senses that the two senses of the sense pair have already been linked to in the (respective) other dictionary, and
    - The total number of senses that the sense pair's lexeme has in either dictionary

Only if the prediction *p* exceeds the threshold *t* for a given sense pair is the sense pair classified as a link; else it is classified as a non-link. Given the sense linker's binary classification as well as the initial prediction *p* and the threshold *t*, we are tasked with calculating the link probability. Calculating this probability is complicated by the two-level structure described above. The issue is that the prediction *p* from the first level does not provide clear cut-offs for when a link is

made. In fact, because of the threshold *t* from the second level, there exist sense pairs that are not linked despite having a higher prediction *p* in the first level than other sense pairs that are linked. For example, a sense pair with an initial prediction *p=0.8* and a threshold *t=0.9* is not linked while a sense pair with an initial prediction *p=0.3* and a threshold *t=0.2* is linked.

4.2 Measuring Certainty

A proper measure of the certainty of prediction must be strictly monotonic regarding the actual certainty of the algorithm (i.e. the predicted link-likelihood). That is, a higher measure directly translates to higher certainty, and a lower measure directly translates to lower certainty. Also, and as a result of the first condition, there must be a clear, numeric cut-off point for creating a binary prediction. For instance, every sense pair with a certainty score equal to or larger than *x* is labelled a link while every sense pair with a certainty score lower than *x* is labelled a non-link (where *x* falls within the range of the certainty score).

The initial prediction *p* meets neither of these conditions because it fails to account for the role of the threshold *t* in deciding the final prediction. Therefore, we chose to quantify the certainty instead by looking at the distance of the initial prediction *p* from the threshold *t*, because this distance directly translates to how *close* a sense pair is to being labelled a link or non-link. Further, this measure satisfies the condition of being directly correlated to the certainty of the algorithm, with a clear cut-off point for the different predictions. We thus have the certainty score *c*:

$$c = p - t$$

Because the ranges of both *p* and *t* are [0, 1], the certainty score *c* ranges from [-1, 1]. Values toward -1 correspond to high certainty of a non-link, values toward 1 correspond to high certainty of a link, and values around 0 correspond to low certainty. Importantly, 0 is the true cut-off for the prediction. That is, a positive score for certainty *c* indicates a predicted link while a negative score indicates a predicted non-link.

After obtaining the estimated certainty value *c*, the next step is to translate this measure to percentage-wise probability estimates for the outcome. That is, beyond a certainty measure, we want the underlying probability that a given sense link is correct.

4.3 Bucket approach

To estimate the link probability of a sense pair, we take advantage of the large amount of annotated data we used for training the predictive algorithm. Using the annotated data, we can compare a new sense pair to annotated sense pairs that are similar. That way, we can estimate the probability outcome for the new sense pair based on the results of the similar sense pairs. Similarity, in this case, is defied across the certainty axis (i.e. the value of *c*). That is, we consider sense pairs to be similar if they received a comparable certainty score *c* by the predictive algorithm.

The certainty variable *c* is divided into several ranges, or buckets, to choose similar sense pairs. To estimate the probability that a given new sense link is correct, we take the relative frequency of true, annotated sense links among all annotated sense pairs in the same respective bucket. For instance, consider a new sense pair with a certainty score of 0.78. If we chose to divide the certainty variable *c* into 20 equidistant (evenly spaced) ranges, or buckets, then the relevant bucket for the new sense pair is the range [0.7, 0.8). Therefore, to estimate the link probability of the new sense pair, we calculate the percentage of true links among the annotated sense pairs with a certainty score *c* between 0.7 and 0.8. This percentage serves as the link probability of the new sense pair.

In our previous work (Saurí et al., 2019), we noted a sizeable variability in data distributions across the different parts of speech, leading us to divide the dictionaries into sense pairs by part of speech. Since this observation also affects the predictions, we again consider each part of speech separately in applying the bucket approach to estimating link probabilities. Figure 1. shows the distribution of the sense pairs regarding the certainty score *c* for each part of speech when dividing *c* into 10 equidistant buckets (evenly spaced ranges). It also visualises the bucket approach: Each bar is a bucket, and the proportion of true links quantifies the link probability for a new sense pair falling into the bucket.

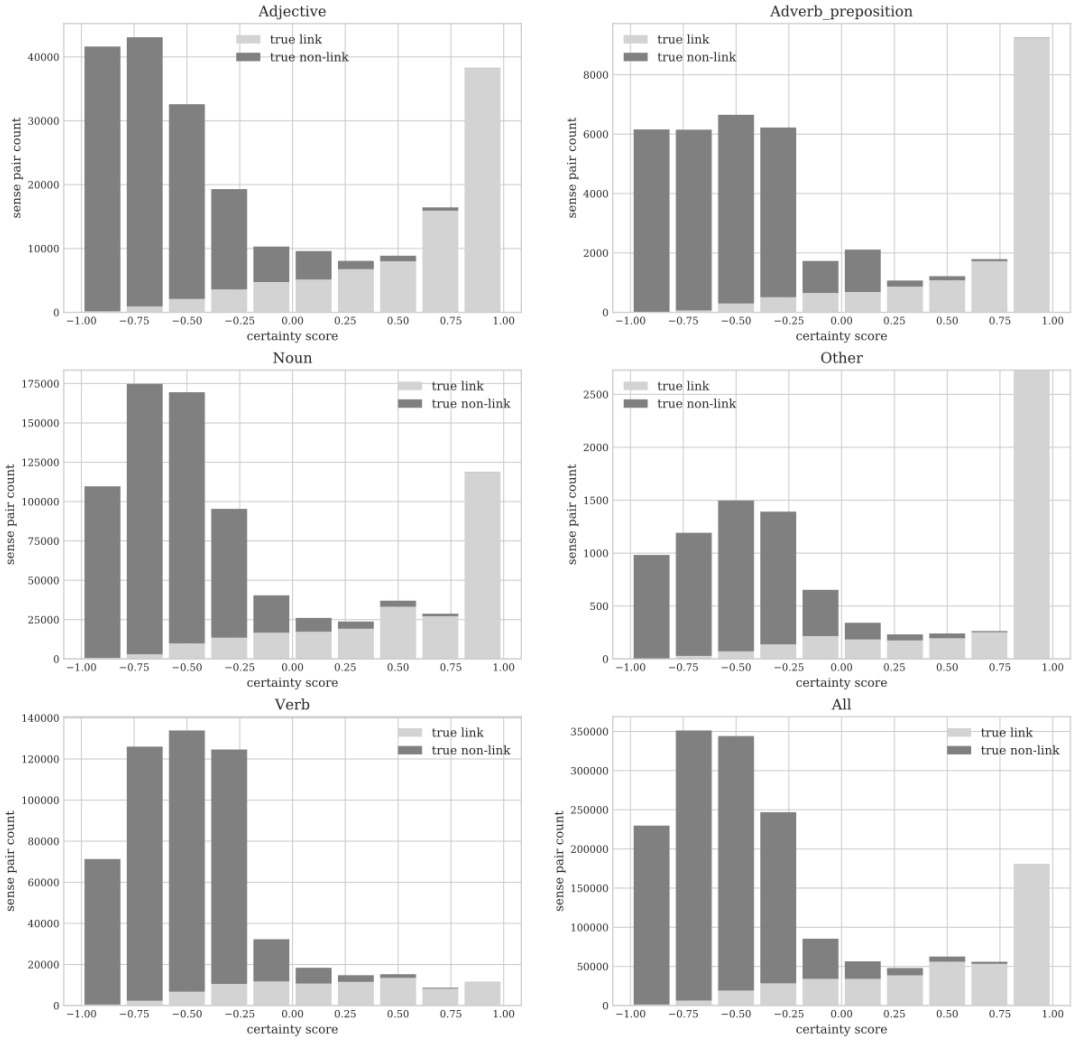

Figure 1: Distribution of sense pairs according to certainty score by part of speech using 10 equidistant buckets

The accuracy of link probabilities estimated with the bucket approach crucially depends on the choice of bucket sizes. When a bucket is too broad, new sense pairs that have considerably unlike certainty scores receive the same probability estimate. The resulting insufficiently fine discrimination of certainty scores is an example of under-fitting. That is, by literally throwing too many sense pairs into the same bucket, we fail to adequately represent the underlying differences in link probabilities that exist in the data. On the flip side, when a bucket is too narrow, the number of annotated sense pairs may be too small, thus rendering the relative frequency measures unrepresentative as well. A scenario exemplifying such kind of overfitting is a sudden, yet random, drop in the link probability for a bucket with a range comprising high certainty scores.

Finding the proper bucket sizes is an empirical matter. Therefore, we experimented with several ways of partitioning the certainty score *c* into buckets. One approach, coined *equidistant buckets*, divides the certainty score *c* into several buckets that each cover a range of the same size. For instance, a division into 20 equidistant buckets means that the first bucket includes all the links with a certainty score *c* in the range [-1, -0.9), the second bucket includes all the links with a certainty score *c* in the range [-0.9, -0.8), and so on to the last bucket covering the range [0.9, 1]. This approach has the advantage of being straightforward and independent of the training data. The width of each bucket is impervious to the distribution of the training data. The disadvantage of the approach is that number of links per bucket can vary hugely between buckets, as some certainty scores may occur more frequently than others. As shown in Figure 1., the distribution of the data is heavily skewed towards extreme certainty scores (close to -1 or close to 1) for every part of speech. This skew has the effect that buckets near the extremes have more sense pairs than the buckets near a certainty score of 0, which biases the relative frequency measures in the different buckets.

The second approach, coined *quantile buckets*, divides the certainty score *c* into several buckets that each contain the same number of sense pairs. For instance, if there are 2000 sense pairs in total and we set the number of buckets to 20, then the quantile approach involves first finding the cut-off points for the buckets across *c* such that each bucket covers 100 sense pairs. These cut-off points then define the extent, or width, of each bucket. As a result, the buckets cover ranges of variable sizes but containing the same number of sense pairs. The advantage of this approach is that there is no risk of

disproportionate buckets since, by definition, every bucket has the same number of links. The main disadvantage is that the determination of the buckets is dependent on the distribution of the training data, which may or may not correspond to the distribution of the data that the approach is applied to.
Figure 2. visualises the two different approaches, with all sense pairs with the part of speech *verb* divided into six equidistant buckets (left) and six quantile buckets (right) for comparison. Note that the bucket width is constant for equidistant buckets and varies for quantile buckets, while the number of sense pairs per bucket varies for equidistant buckets and is constant for quantile buckets.

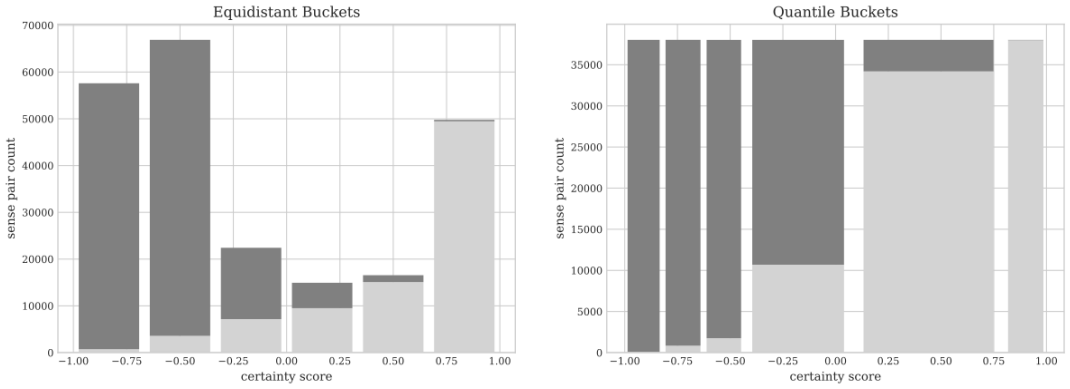

Figure 2: Comparison of equidistant buckets and quantile buckets using the example of *verb* sense pairs divided into six buckets

We further experimented with the absolute number of buckets. For every part of speech, we tried out 10, 50, 100, 500 and 1000 equidistant and quantile buckets.

4.4 Validation

The 'goodness of fit' of each bucket approach variety was quantified by calculating the logistic likelihood ratio (with base 10) for the resulting sense link probabilities:

$$LLR_{10} = N \log_{10} 2 + \sum_{1} \log_{10} p_i + \sum_{0} \log_{10}(1 - p_i)$$

N is the number of sense pairs in the entire dataset, and *p* is the predicted link probability for a given sense pair. The summations are for all true links (1) and all true non-links (0) respectively. Due to a large number of sense pairs, the resulting likelihood ratios blow out of proportion, which is why we additionally calculated the average likelihood ratio for one sense pair as follows:

$$avgLR = 10^{LLR_{10}/N}$$

avgLR ranges from (0, 2] and represents how much better or worse the predictions are (on average) compared to chance. Chance is defined as a 50-50 guess, a flip of a fair coin, to make the prediction. An *avgLR* score of 1 implies that the probability estimates are no better (or worse) than chance. Optimal predictions, where every link is predicted with $p = 1$ and every non-link is predicted with $p = 0$, result in an *avgLR* score of 2, meaning that, for each sense pair, the prediction was twice as good as chance. This upper-bound is logical, as pure chance classifies a sense pair correctly with probability $p = 0.5$ and we cannot do better than probability $p = 2 * 0.5 = 1$.
Random guessing is a low bar to compare our probability estimates against, especially since the algorithm in Saurí et al. (2019) does the heavy lifting of prediction. Therefore, we also compare the results against a well-informed baseline. The baseline is calculated by taking the true positive rate of all predicted links in the training data as the probability *p* for all predicted links, and the false negative rate of all predicted non-links in the training data as the probability *p* for all predicted non-links. For instance, if in the training data 84% of all predicted links were true links and 12% of all predicted non-links were true links, then the link probability estimates are 0.84 for predicted links and 0.12 for predicted non-links. Note that we can expect the baseline to perform much better than chance, as it takes into account the prediction of the algorithm. That is, the baseline probabilities are informed by what the algorithm has learned about sense pairs and sense links. Note further that this baseline is equivalent to the bucket approach with two equidistant buckets (one for predicted links and one for predicted non-links), and therefore can be seen as the most basic version of the approach proposed in this paper.
The bucket approach was cross-validated using each of the six dictionaries used in the training of the binary classifier from Saurí et al. (2019) as one fold. That is, five dictionaries were used to create the buckets and calculate the relative frequencies of true links. These relative frequencies were then used to estimate the link probability of every sense pair in

the sixth dictionary. For each part of speech separately, we identified the best performing bucket approach as the one that had the highest mean *avgLR* score across the six dictionaries. Table 1. summarises the cross-validation results, showing the baseline as well as the best performing bucket approach for each part of speech.

Part of speech	Adjective	Adverb / Preposition	Noun	Other	Verb
Baseline (*avgLR*)	1.543	1.605	1.530	1.524	1.531
Best buckets (*avgLR*)	**1.656**	**1.710**	**1.638**	**1.634**	**1.614**
Best bucket approach	500 equidistant	100 equidistant	100 equidistant	50 quantile	500 equidistant

Table 1: Average Likelihood Ratio (*avgLR*) for the baseline and the best performing bucket approach for each part of speech

As expected, the baseline performs reasonably well, predicting a sense pair on average around 1.5-1.6 times better than chance. For every part of speech, the best performing bucket approach outperforms the baseline, with an *avgLR* score consistently higher at around 1.6-1.7. Thus, cross-validation confirms that the bucket approach does enable us to generate more precise probability estimates for each sense pair individually. Quantile buckets emerged as the best approach only for the part of speech *other*, and equidistant buckets for all other parts of speech. The optimal number of buckets ranges from 50 to 500, plausibly with some correlation to the size of the dataset (e.g. *verb* has many more sense pairs than *other*).

5 Results

The sense linker developed in Saurí et al. (2019) in conjunction with the algorithm presented in this paper were used to link an English-Arabic bilingual dictionary and a Portuguese-English bilingual dictionary to a monolingual English dictionary. For both Arabic links and Portuguese links, we took a random sample of around 1000 sense pairs for each part of speech and obtained the gold standard from expert lexicographers. Table 2. and Table 3. show the results of the external validation using the random samples compared to the baseline for Arabic and Portuguese links, respectively. Recall that an *avgRL* score of 1 represents performance equivalent to chance, while an *avgRL* score of 2 corresponds to perfect predictions (i.e. links with a probability of $p = 1$ and non-links with a probability of $p = 0$). The baseline, again, is equivalent to the bucket approach with two equidistant buckets.

Part of speech	Adjective	Adverb / Preposition	Noun	Other	Verb
Baseline (*avgLR*)	**1.419**	1.454	1.557	1.154	1.545
Best buckets (*avgLR*)	1.319	**1.500**	**1.584**	**1.260**	**1.598**
Best bucket approach	500 equidistant	100 equidistant	100 equidistant	50 quantile	500 equidistant

Table 2: Average Likelihood Ratio (*avgLR*) results compared to the baseline for a random sample of Arabic links

Part of speech	Adjective	Adverb / Preposition	Noun	Other	Verb
Baseline (*avgLR*)	1.383	1.477	1.470	1.272	1.479
Best buckets (*avgLR*)	**1.437**	**1.528**	**1.563**	**1.304**	**1.537**
Best bucket approach	500 equidistant	100 equidistant	100 equidistant	50 quantile	500 equidistant

Table 3: Average Likelihood Ratio (*avgLR*) results compared to the baseline for a random sample of Portuguese links

The improvements over the baseline are more modest than was the case with cross-validation. This result was to be expected since the optimal bucket approach was chosen during cross-validation, while only the bucket approach identified during cross-validation was used for external validation. For the links of the Portuguese-English bilingual dictionary, the previously determined optimal bucket approach consistently outperforms the baseline, albeit to a lesser extent than during cross-validation, with a mean *avgLR* score improvement close to 0.05 across the different parts of

speech. For the links of the Arabic-English bilingual dictionary, we obtain similar results for four out of the five parts of speech. However, for adjectives, the baseline performs better than the previously determined optimal bucket approach on the random sample. Potential reasons are addressed in the Discussion section.

6 Discussion

6.1 Interpretation of Results

The results show that, using the bucket approach described in this paper, we can obtain sense link probability estimates for individual sense pairs that outperform more general approaches like the baseline. The part of speech group *other* proves most challenging to predict. A potential reason may be that *other* subsumes many different parts of speech (e.g. conjunctions, pronouns, determiners), which may not generalise well to different languages. Another possible explanation is simply that *other* has the fewest samples (i.e. sense pairs) in the training data, thus making the relative frequency measures used to estimate probabilities less reliable. Another noteworthy observation is that the baseline performs better than the more nuanced bucket approach (only) for the part of speech group *adjective* in the English-Arabic dictionary. This exception suggests that the probability estimates for Arabic adjectives were overly optimistic (as false predictions with high certainty are punished disproportionately, strongly affecting the outcome), possibly because the features that are highly indicative of sense links for adjectives in general are less relevant for Arabic adjectives in specific. Of course, given the nature of using random samples for validation, fluctuations in performance may also, to a certain degree, be explained by randomness.

6.2 Implications for Automated Sense Linking

To make use of automated means of sense linking in practice, the ability to quantify the quality of the linked data is paramount. Most approaches to automated sense linking rely heavily on general quality estimates for entire datasets. Such estimates are often unreliable, as they do not generalise well to other datasets. One example proving this notion comes from the stratification by part of speech. As shown in Figure 1. in the Method section, and likely due to high polysemy, *verb* sense links are much harder to predict than those of other parts of speech. Accordingly, quality estimates for all parts of speech are not representative of, for instance, the quality of a linked dataset consisting only of verbs.

Some automated sense linking approaches attempt to make more nuanced statements regarding link quality by introducing thresholds along the certainty axis. They may say, for instance, that sense pairs above one certainty threshold have one estimated precision, while sense pairs above another certainty threshold have different estimated precision. The issue here is the implicit assumption that the distribution of certainty scores remains the same across different datasets. Using the example of *verb* sense pairs again, we can see that this is a false assumption: *Adjective* sense pairs with a certainty score $c > 0.25$, for instance, have higher precision than *verb* sense pairs with a certainty score $c > 0.25$ because there are many more *adjective* sense pairs with a certainty score c close to 1 than *verb* sense pairs.

In other words, general quality statements are not sufficient as they ignore the differences in distribution across different datasets. Probability estimates, on the other hand, circumvent this issue as they provide estimates for each sense pair independently, thereby being impartial to the distribution of the larger dataset. A further benefit of generating probability estimates for individual sense pairs is that quality estimation on the sense pair level opens up the opportunity of integrating automated sense linking in lexicographic processes without imprudently relinquishing the expertise of lexicographers. By linking sense pairs with high link-probability automatically while manually reviewing sense pairs with high uncertainty, it is possible to improve the quality of automatically generated sense links with limited editorial resources. Quality estimation, therefore, must not be seen merely as a validation mechanism for a given approach, but rather as a tool for making possible high-quality, (semi-)automated sense linking.

Probability estimates provide substantial benefits that help drive the implementation of automated sense linking in the applied, industrial context.

6.3 Implications for Lexicography

The work presented also has important implications for lexicography more generally. As mentioned above, the ability to quantify the probability that an automatically linked sense pair is indeed a link allows us to merge automated and manual sense linking. The sense pairs that can be automatically linked with high probability can be considered links without further inspection, while sense pairs automatically linked with low certainty (i.e. a probability near 0.5 for the binary task of sense linking) can be passed on to lexicographers for expert validation. This method can significantly speed up the process of linking lexical resources while retaining control over the quality of links. Precise probabilities also enable us to quantify the trade-off between the quality of the links and the amount of work needed from lexicographers. In that way, lexicographers can gain direct insights into how their work improves the quality of automatically linked lexical data, and base decision-making on the quality estimates.

Additionally, probabilistic automatic sense linking can point lexicographers to differences, incongruities, and omissions across lexical resources. For instance, when a sense in one resource has no corresponding sense in another resource (i.e. all relevant link probabilities are small), we may infer that in the latter resource the sense is (correctly or mistakenly) excluded or subsumed in another sense. Inversely, when a sense in one resource is linked with high probability to several senses in another resource, we have reason to believe that the latter resource applies a finer level of granularity. Investigating such cases can be useful for lexicographers when revising and expanding (linked) lexical resources.

Ultimately, lexicographers provide the gold standard for supervised machine learning based approaches to automated sense linking. As such, all linked data annotated by lexicographers is useful in improving the algorithms' outcomes by

providing more training data. Furthermore, inspecting sense pairs of interest may provide insights into how to improve a given algorithm or, in the very least, point out its shortcomings. Sense pairs of special interest may be false positive predictions that have a high probability score, reversely false negatives that have a low probability score, and sense pairs with high uncertainty.

7 Conclusion

This paper has described a method for obtaining probability estimates for predicted sense links between two dictionaries. These probabilities play an essential role not only in reliably estimating the quality of a given set of sense links but also in differentiating sense links with high certainty from those with low certainty. This differentiation makes possible the deliberate trade-off between the correctness and the completeness of generated links, as well as the optimal improvement of link quality through limited editorial input by lexicographers. Both reliable quality estimation and semi-automated sense linking are vital points in moving automated sense linking from a research interest to a content production tool. As part of the broader efforts of linking lexical content at Oxford University Press, the presented quality estimation algorithm achieves precisely that.

8 References

Alper, M. (2017). Auto-generating Bilingual Dictionaries: Results of the TIAD-2017 Shared Task Baseline Algorithm. In *Proceedings of the LDK 2017 Workshops, co-located with the 1st Conference on Language, Data and Knowledge*, pp. 85–93.

Arcan, M., Torregrosa, D., Ahmadi, S., & McCrae, J. P. (2019). Inferring translation candidates for multilingual dictionary generation with multi-way neural machine translation. Paper presented at the *Translation Inference Across Dictionaries Workshop (TIAD 2019), Leipzig, Germany, 20-23 May*, doi:10.13025/S89K9J

Donandt, K., Chiarcos, C., & Ionov, M. (2017). Using Machine Learning for Translation Inference Across Dictionaries. In *Proceedings of the LDK 2017 Workshops*.

Gollins, T., & Sanderson, M. (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. ACM.

Gracia, J., Kabashi, B., & Kernerman, I. (2019) *Proceedings of TIAD-2019 Shared Task – Translation Inference Across Dictionaries*. Co-located with the 2nd Language, Data and Knowledge Conference (LDK 2019) Leipzig, Germany, May 20, 2019. CEUR Workshop Proceedings, Vol. 2493. http://ceur-ws.org/Vol-2493/

Kaji, H., Tamamura, S., & Erdenebat, D. (2008). Automatic Construction of a Japanese-Chinese Dictionary via English. In *LREC 2008*.

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

Mausam, Soderland, S., Etzioni, O., Weld, D., Skinner, M., & Bilmes J. (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. ACL.

Niculescu-Mizil, A., & Caruana, R. (2005). Obtaining calibrated probabilities from boosting. In Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence (UAI'05). AUAI Press, Arlington, Virginia, USA, pp. 413–420.

Proisl, T., Heinrich, P., Evert, S., & Kabashi, B. (2017). Translation Inference across Dictionaries via a Combination of Graph-based Methods and Co-occurrence Stats. In LDK Workshops

Saurí, R., Mahon, L., Russo, I., & Bitinis, M. (2019). Cross-Dictionary Linking at Sense Level with a Double-Layer Classifier. LDK.

Shezaf, D., Rappoport, A. (2010). Bilingual lexicon generation using non-aligned signatures. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL '10). Association for Computational Linguistics*, USA, pp. 98–107.

Tanaka, K., & Umemura, K. (1994). Construction of a Bilingual Dictionary Intermediated by a Third Language. In *Proceedings of COLING'94*, pp. 297–303, 1994.

Varga, I., & Yokoyama, S. (2009). Bilingual dictionary generation for low-resourced language pairs. In *Proceedings of EMNLP*, pages 862–870. URL: http://www.aclweb.org/anthology/ D09-1090.

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

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