More is not always better: balancing sense distributions for all-words Word Sense Disambiguation

Marten Postma, Ruben Izquierdo Bevia, Piek Vossen
Vrije Universiteit Amsterdam
De Boelelaan 1105, 1081 HV Amsterdam
The Netherlands
m.c.postma@vu.nl, rubensanvi@gmail.com, piek.vossen@vu.nl

Abstract

Current Word Sense Disambiguation systems show an extremely poor performance on low frequent senses, which is mainly caused by the difference in sense distributions between training and test data. The main focus in tackling this problem has been on acquiring more data or selecting a single predominant sense and not necessarily on the meta properties of the data itself. We demonstrate that these properties, such as the volume, provenance, and balancing, play an important role with respect to system performance. In this paper, we describe a set of experiments to analyze these meta properties in the framework of a state-of-the-art WSD system when evaluated on the SemEval-2013 English all-words dataset. We show that volume and provenance are indeed important, but that approximating the perfect balancing of the selected training data leads to an improvement of 21 points and exceeds state-of-the-art systems by 14 points while using only simple features. We therefore conclude that unsupervised acquisition of training data should be guided by strategies aimed at matching meta properties.

1 Introduction

The task of automatically selecting the meaning of a word in a linguistic context, known as Word Sense Disambiguation (WSD), has been and is one of the main challenges for NLP. Despite the huge amount of research that has been carried out to analyze and tackle this problem, it is nevertheless still considered unsolved. The best performances on the SemEval-2013 task 12: "Multilingual Word Sense Disambiguation" English all-words subtask (from now on *sem2013-aw*) (Navigli et al., 2013) was 72.28 (Weissenborn et al., 2015)) out of competition and 64.7 in competition (Gutiérrez et al., 2013), exceeding the naive Most Frequent Sense (MFS) baseline by only 10 points at most.

Supervised machine learning systems have been successful for WSD and it is commonly believed that the problem of the low performance can be solved by providing more (manually annotated) training data. However, in addition to the sparseness of training data, another aspect of the problem is the Zipfian distribution of word senses (McCarthy et al., 2007). In both training data and test data, the same single sense of a word tends to heavily dominate, making the MFS not only a baseline that is difficult to beat (Kilgarriff, 2004), but also being used as a fall-back option by many systems.

Given the Zipfian distribution of word senses, it comes as no surprise that WSD systems have a bias towards assigning the MFS (Preiss, 2006). Building upon this observation, Postma et al. (2016) analyzed systems participating in previous Senseval and SemEval competitions with respect to their performance on the MFS and on all other senses, which are called the less frequent senses (LFS). This study convincingly showed that systems excel at identifying MFS instances, but that all systems underperform on LFS instances. For instance on the sem2013-aw task, systems performed on average around 80% in accuracy on the MFS instances, but they hardly achieved on average 20% accuracy for the LFS instances.

The main challenge for WSD is therefore not only to acquire more training data, but to also address the overfitting towards the majority case in order to boost the performance for the LFS cases. Assuming that

Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 3496–3506, Osaka, Japan, December 11-17 2016.

their low performance is due to both the sparseness of training data and differences in sense distributions, the challenge is how to acquire more training data and obtaining the right distribution. In this paper, we therefore investigate the following research questions:

1. **Volume**: What is the influence of using more training data without distinguishing between MFS and LFS cases?
2. **LFS**: What is the influence of only adding more LFS training examples?
3. **Provenance**: What is the effect of training using manually annotated data versus automatically annotated data?
4. **Balancing**: What is the effect of mimicking the perfect target distribution in the training data?

By providing evidence for each research question, we firstly demonstrate that more data has a small impact on the performance of a state-of-the-art supervised system (It Makes Sense (IMS), (Zhong and Ng, 2010)). Interestingly enough, the type of data has a bigger impact, more specifically, adding silver data with a better fit to the test set with respect to time and genre appears to be better than adding more manually annotated data. Our biggest contribution is however that a perfect balance of data has a major impact on the performance. Balancing the training data according to the sense distribution of the test data boosts the results for the LFS cases while maintaining the high performance for the MFS instances. Following this assumption, our experiments presented in this paper reach an overall accuracy of 86.8 as an upper ceiling. This points towards the conclusion that the evaluation data sets have so-called "long-tail details" that need to be modeled to obtain a high-performance in addition to the properties of the head of the distribution. We therefore conclude that the distributional effect is more complex than suggested in (McCarthy et al., 2004a), who focus only on the predominant sense, but also has a larger potential if acquisition is guided by strategies to match meta properties.

The paper is structured as follows: in Section 2 we discuss related work, following by a description of the resources and the evaluation framework (Section 3). The experiments and the results are presented in Section 4. Finally, we discuss the results in Section 5 and conclude the paper in Section 6. All the training data, system output files and scripts required to fully reproduce the experiments presented in this paper have been made publicly available at: `https://github.com/cltl/MoreIsNotAlwaysBetter`.

2 Related Work

Many natural phenomena can be described by power laws (Newman, 2005), ranging from city populations to name frequencies. Word sense distributions are no exception to this interesting phenomenon and also show Zipfian characteristics (McCarthy et al., 2007; Kilgarriff, 2004). This means that, given any document or collection of documents, one sense of a lemma is usually overrepresented, while the other senses are barely used. The MFS consequently plays a critical role in the task of WSD and establishes a challenge for systems. In evaluation, the MFS is usually used as a hard to beat baseline and it is therefore also used as a fallback strategy by systems.

In general, WSD systems perform well on the MFS instances, whereas their performance drops dramatically for the LFS instances (Postma et al., 2016). This is not surprising considering the Zipfian distribution of senses which has a big impact on supervised approaches. Given the dominance of MFS examples in the training data, this results in better sense representations for these cases. This unbalance contributes to the system bias towards the MFS. For unsupervised approaches, in particular graph-based approaches, the MFS of a lemma also appears to have a higher connectivity in the graph (Calvo and Gelbukh, 2015), hence also favoring it in the sense assignment phase of a WSD system. Overall, less attention has been paid to the less represented and less frequent senses, despite the fact that these provide the biggest room for improvement (Postma et al., 2016).

Although the MFS in the training data often coincides with the MFS of the task, this is not always the case, specially in cross-domain or genre scenarios. A system able to detect the predominant sense of a lemma within a target document would obtain a vast increase in accuracy. McCarthy et al. (2004b) and Koeling et al. (2005) provided the first proof of concepts in order to demonstrate this. They collected for

each target word k nearest neighbors using distributional similarity, where the similarity between each sense of the target word and the nearest neighbors is determined by WordNet similarity measures. The sense with the highest similarity is chosen as the predominant sense. Building upon the same idea, Chan and Ng (2005) apply two algorithms, a confusion matrix algorithm and an Expectation-maximization-based algorithm, to determine the sense distribution of the test data. This information is fed into the systems to improve the sense assignment.

Similar approaches have been used to tackle the task of Word Sense Induction. The work presented in Lau et al. (2012) introduces a Topic Modeling approach based on LDA (Blei et al., 2003) and HDP (Teh et al., 2012) for deriving clusters that can be identified for different senses of a word. As an intermediate outcome, the authors also provide the expected predominant senses in the target corpus. Similarly, Boyd-Graber and Blei (2007) apply a model that considers the words of a document generated coherently to the topic distribution of that document. The most likely sense for each instance of a word is predicted from this topic distribution considering the words in the context. Finally, Lau et al. (2014; 2012) focused on the detection of novel senses, which might be considered as an extreme case of unbalanced acquisition with respect to training and evaluation distributions.

Mismatches between the training and test data have been of central interest to research on domain adaptation (Daume III, 2007; Carpuat et al., 2013; Jiang and Zhai, 2007). The 2010 edition of the SemEval series (SemEval-2010) proposed a task called "All-words Word Sense Disambiguation on a Specific Domain" (Agirre et al., 2010). The aim was to analyze to what extent WSD systems are sensitive to specific domains when there is no annotated data available for that domain. In the same direction, another goal was to investigate how a general domain WSD system should be adapted to perform properly when the sense distribution is unknown and different to the sense distribution of traditional corpora used for training machine learning models. This evaluation showed that the most successful approaches included knowledge from the specific domain in different forms. For example, Kouno et al. (2015) present a framework for WSD where an unsupervised approach is applied to abstract the features across different domains and then feeds these features into a Support Vector Machine. A semi-supervised framework is introduced by Agirre and Lopez de Lacalle (2008). Several domains are used to establish cross-domain experiments, where two supervised machine learning WSD systems (k-NN and SVM) are applied on one domain and evaluated on a different one. Singular Value Decomposition (SVD) is employed to find correlations between terms, alleviate the scarcity of data, and extract examples from unlabeled data. The authors show an improvement on the cross-domain setting when including the SVD technique to add training data of the target domain.

The importance of the MFS bias in training data was already highlighted in previous work. For instance, Agirre and Martinez (2004) try to overcome the problem of this skewness by automatically acquiring examples from the Internet using an heuristic based on monosemous words. Improvement is achieved on the Senseval-2 task for nouns with less than 10 examples in SemCor (Miller et al., 1993).

Finally, researchers combined the task of WSD with Entity Linking to improve the performance on the disambiguation part. For example, Weissenborn et al. (2015) jointly disambiguate nouns and entities by exploiting the links between them in BabelNet (Navigli and Ponzetto, 2012). In addition, the Babelfy system (Moro et al., 2014) goes one step further by also making use of interlingual relations.

Despite these efforts to either acquire more data or adapt the distributions, none of these systems gained an improvement big enough to consider the problem as solved. Our work differs from these approaches mainly in that we analyze more precisely the contribution of the volume, nature, and distribution of the training data in relation to the properties of the test data. The experiments are designed to isolate each phenomenon and be able to extract meaningful conclusions. For example, whereas Agirre and Martinez (2004) acquire more data for all senses of low frequent words, we focus on adding more examples for the less frequent senses instead. Just by obtaining more examples for a word does not imply that we are able to create better sense representations for all the senses of this word. Similarly, whereas McCarthy et al. (2004b) and Koeling et al. (2005) restrict the system to find a deterministic predominant sense, we propose that the probabilities of the target data need to be used to obtain a more fine-grained behavior. We show that properly balancing the probabilities of the acquired data gives a

major improvement even using a straight-forward machine learning system with a basic set of features.

3 Methodology

In this section we describe our training and evaluation framework: which WSD system has been selected, what corpora have been used to create our models and how these models have been evaluated. We have set up the framework in such a way that we can systematically vary the type of training data used, the volume of training data and the sense distribution in this data. The impact of the variables is then tested using the same state-of-the-art supervised WSD system. We measure the overall performance in addition to the performance on the MFS and LFS cases and also generate some statistics to highlight the characteristics of the training data.

3.1 Training data sets

In our experiments, we exploit three sources of English sense annotated data: SemCor, Princeton WordNet Gloss Corpus, and what we have called "Wordnet2Wikipedia". The last corpus was created as part of the experiments and hence its creation will be described in more detail below.

Semcor (SC) The Semantic Concordance (Miller et al., 1993), or SemCor (SC), is a corpus containing approximately 240,000 sense annotated words. The tagged documents originate from the Brown corpus (Francis and Kucera, 1979) and cover various genres. The corpus contains annotations for more than 20,000 lemmas. The creators note that the main focus when creating the corpus was on word frequencies, and not on sense frequencies. This might explain why the proportion of the MFS on the total amount of annotated senses in this corpus is (unnaturally) high: more than 70%.

Princeton WordNet Gloss Corpus (GC) The Princeton WordNet Gloss Corpus, or GC, is a special sense annotated corpus. It does not contain annotations of natural text in documents but of the WordNet glosses (Fellbaum, 1998). The corpus contains more than 310,000 annotated words, which is significantly more than in SemCor. Annotations exist for approximately 15,000 lemmas, which is less than in SemCor. The MFS proportion is around 55%, which makes sense given that glosses are tagged and not natural text.

Wordnet2Wikipedia (WW) For the last source of sense annotated data, WordNet2Wikipedia, or WW, we exploit the existing relation between WordNet and Wikipedia as present in BabelNet 2.5 (Navigli and Ponzetto, 2012). BabelNet 2.5 contains the relations *direct* and *redirect*, which link a WordNet sense to a Wikipedia entry. This relation only exists for nouns in the resource. For each target sense of a target lemma: 1. we check whether a link to Wikipedia exists for this target sense. 2. if so, we extract all sentences from the Wikipedia article with the target lemma and tag the target lemmas with the target sense. By exploiting this relation in BabelNet, we are able to extract 43,000 training examples for the 751 lemmas of the sem2013-aw competition. In addition, on average, 63% of the examples we extract are examples of LFS instances.

3.2 The WSD System

We used the "It makes sense" (IMS) WSD system in our experiments (Zhong and Ng, 2010). There are several reasons for choosing IMS. First of all, IMS has shown to achieve a very high performance in the Senseval and SemEval all-words tasks. Secondly, it is an open source system that provides a flexible framework, allowing us to train and evaluate it with different kinds and amounts of data. Finally, IMS is based on a Support Vector Machine (SVM) learning engine, which is a linear classifier known to suffer not as much from unbalanced training as other learning paradigms (e.g. probabilistic learning paradigms like Naive Bayes).[1]

IMS uses three features in its default setting: surrounding words, part-of-speech tags, and collocations in a certain window around the target word. We used the system with the default setting in terms of features and learning parameters. IMS creates word experts, which means that one single classifier is

[1] http://www.win-vector.com/blog/2015/02/does-balancing-classes-improve-classifier-performance

built for a specific lemma (with the corresponding part-of-speech tag). We could easily adapt the specific word experts by manipulating the input training data while keeping all the other settings the same.[2]

3.3 Evaluation framework

For evaluation, we selected the test set from SemEval-2013, task 12: Multilingual Word Sense Disambiguation (**sem2013-aw**, (Navigli et al., 2013)). The task consisted of two disambiguation tasks: Entity Linking and Word Sense Disambiguation for English, German, French, Italian, and Spanish. We focused on the WSD part using WordNet as a sense repository. The test set contains 13 articles obtained from previous editions of the workshop on Statistical Machine Translation.[3] The articles cover different domains, ranging from sports to financial news. With respect to the English WSD part, there are 1,644 test instances in total, all nouns. The sense repository used to tag these instances is WordNet version 3.0. Based on this sense repository, we computed the number of instances annotated with the MFS, and the number of instances annotated with one of the LFS. Out of the 1,644 test instances, the MFS applies in 1,035 cases, while one of the LFS applies in the rest of 609 cases. This gives a baseline of 62.96% for the MFS heuristic.

3.4 Experiments

In order to gain insight into system performance with respect to the quality, quantity, and distribution of the training data, four main research questions have been formulated, which we will repeat here for the sake of clarity:

1. **Volume**: What is the influence of using more training data without distinguishing between MFS and LFS cases?
2. **LFS**: What is the influence of only adding more LFS training examples?
3. **Provenance**: What is the effect of training using manually annotated data versus automatically annotated data?
4. **Balancing**: What is the effect of mimicking the perfect target distribution in the training data?

We use several selection techniques to test each research question. We distinguish between a *Base* corpus from which we add all training instances available, and an *Expansion* corpus, from where subsets of instances are extracted by applying different selection techniques. SemCor (**SC**) is used as a Base and the WordNet gloss-corpus (**GC**) and the WordNet-Wikipedia corpus (**WW**) as Expansions, representing both more Volume and different types of annotation (Provenance). In addition, we defined the following selection techniques when expanding the training data:

All: all instances, both MFS and LFS, are added. This technique will allow us to provide evidence for the **Volume** and **Provenance** research questions.

LFS: only the LFS instances are added, which provides insight into the **MFS** versus **LFS** research question.

Top-down: to provide evidence for this question, we use the sense distribution of the test data starting from the MFS. For every lemma in the test set, we attempt to fit the training distribution as much as possible to the test sense distribution. In this *top-down* approach, we start with the sense with the highest relative frequency in the test set and determine the number of examples in our experiment for this sense. We then calculate the number of examples for the other senses relative to the number of examples for the sense with the highest relative frequency. When a sense does not occur in the test set, we assign to this sense a default number of 1 (top-down-1) or 5 (top-down-5) training examples, which makes sure that we perform balancing instead of filtering. Without the assignment of the default number of senses for missing senses in the test data, we would on average reduce the average polysemy to almost one, which would make the task almost solved.

[2]During our experiments, we noticed that IMS produces exceptions when using larger training sets than SemCor. We solved this by including a number of checks in the source code and by modifying two files with respect to the original IMS distribution. These files are included under the folder *ims_amended_files* in the data and scripts package delivered together with the paper. The installation script will take care of copying them to the proper place in the IMS project structure and recompile the full library. The changes have been documented in the README file.

[3]`http://www.statmt.org`

Bottom-up: to provide evidence for this question, we use the sense distribution of the test data starting from the LFS. This approach is almost identical to the *Top-down* approach but now we use the number of examples for the sense of a lemma with the lowest relative frequency in the test set as a starting point. When a sense does not occur in the test set, we assign a default number of 1 (Bottom-up-1) or 5 (Bottom-up-5) examples to the sense. Both the **Bottom-up** technique and the **Top-down** technique provide evidence for the **Balancing** research question.

To allow replication of our results and stimulate further research, all the data and scripts for training and evaluating the models have been made publicly available. There is an installation script to take care of the downloading and installation of the required libraries and data sets. There is also one main script that runs all the experiments described in Section 4. The package can be found here: `https://github.com/cltl/MoreIsNotAlwaysBetter`.

4 Results

Table 1 presents the results for all our experiments. Each experiment has been defined to analyze one of our research questions and is hence based on a certain training dataset, which is the result of applying various selection techniques, which include the type of data, the volume, and the sense distribution. For each run of an experiment (**ID**), we present the Base corpora (**Base**), the Expansion Corpora (**Expansion**), the selection technique used on the Expansion corpora (**Technique**), the overall accuracy (**Acc**), the accuracy on the MFS instances (Acc_{mfs}), the accuracy on the LFS instances (Acc_{mfs}), the micro average between the accuracy on the MFS and LFS instances (**MicroAV**), the number of training instances (**#**), the average number of training instances per lemma (**Avg#**), and the average percentage of MFS instances per lemma (Dom_{mfs}). The last three values provide information about the quantitative characteristics of the training data.

				WSD results				**Training stats**		
ID	Base	Expansion	Technique	Acc	Acc_{mfs}	Acc_{lfs}	MicroAV	#	Avg#	Dom_{mfs}
1	SC	-	-	65.60	**95.60**	14.80	55.20	22k	43	70
2	SC	GC	All	66.80	89.10	29.10	59.10	60k	110	59
3	SC	GC+WW	All	68.90	90.30	**32.50**	**61.40**	102k	187	52
4	SC	WW	All	**69.30**	**92.80**	29.40	61.10	65k	120	54
5	SC	GC	LFS	63.20	75.70	41.90	58.80	38k	71	43
6	SC	GC+WW	LFS	62.00	70.50	**47.60**	59.05	65k	120	31
7	SC	WW	LFS	**67.50**	**87.50**	33.30	**60.40**	49k	91	44
8	-	SC+GC+WW	Bottom-up1	85.40	95.90	67.50	81.70	46k	87	56
9	-	SC+GC+WW	Bottom-up5	80.40	93.30	58.50	75.90	50k	96	53
10	-	SC+GC+WW	Top-down1	**86.80**	**96.50**	**70.30**	**83.40**	45k	85	55
11	-	SC+GC+WW	Top-down5	82.00	94.40	60.90	77.65	65k	120	54

Table 1: Results of our experiments on SemEval-2013 dataset

Volume and Provenance Experiments *1*, *2*, *3*, and *4* provide insight into the Volume and Provenance research questions. We observe a clear trend that more training data indeed improves the performance. By only adding the GC corpus, the accuracy improves by 1.3 points, and if we add both GC and WW by 3.4 points. These experiments are not solely increasing the number of training examples, they also lower the MFS dominance per lemma as shown in the last column (SC scoring 70, and the expansions scoring lower 59, 42, and 54, respectively) and thus change the balancing or distribution of the senses in the training data. Both factors seem to contribute to the improved overall accuracy. Surprisingly, the best result is not found by using all of the available sense annotated data, as shown by experiment 4. By only adding the WW corpus to SC, we improve the baseline by 3.7 points despite the fact that the examples are extracted automatically (silver) compared to 1.2 points gain when adding an equal amount of data from the manual annotation of GC to SC. We suspect that the WW corpus is more similar to the test data with respect to creation time and genre since they come from Wikipedia whereas the SemCor texts go back to 1961. The GC corpus contains usage examples and definitions, which are older than the test data and timeless while also being more formal in style, written for a different purpose. Finally,

the experiments 2, 3, and 4 all dramatically improve the performance on the LFS instances by more than 15(!) points compared to the baseline in experiment 1. The price for this can be found in a slight drop on the MFS instances. In order to get a better understanding of the improvement of our best run (experiment 4), compared to the baseline, we computed the accuracy per sense rank class, as shown in Figure 1. We can observe a clear improvement of the sense rank classes 2, 3, 4, and 10+ among the LFS cases.

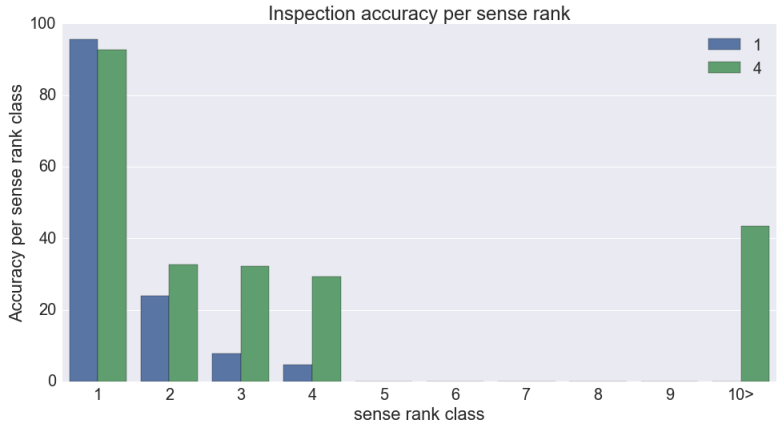

Figure 1: Performance of run 1 (left column) and 4 (right column) with respect to the performance per sense rank class.

LFS What is the effect of only adding LFS instances? We know that some senses of WordNet annotated in the test set are not annotated in SemCor, so they are not available for the training of the IMS system. What will happen if senses with few or no training data get boosted? Experiments 5 to 7 provide evidence to answer these questions. We note that the dominance of the MFS in the training instances now drops to 30-40%. On the other hand, this also results in very impressive performances on the LFS instances, between 33.33% and 47.62%. What we gain on one side, we lose on the other and the only experiment that is able to beat the baseline is run 7, which beats the baseline by 1.9 points.

Balancing Apparently the trick is to find the proper balancing between the MFS and LFS to maintain the high performance of the former and still gain in the performance of the latter. To demonstrate the potential of proper balancing, we apply the Top-down and Bottom-up balancing strategies on the training data using the test data as an approximation of the perfect balance. This can be seen as approximating the upper bound for systems being aware of the sense distribution of the test data. We see in experiments 8, 9, 10, and 11 that properly balancing the distribution gives us the highest gain (21 points) up to an overall accuracy of even 86.8. We see that it enormously boosts the accuracy of the LFS to levels normally achieved for the MFS and it also improves the accuracy of the MFS to 96.52. Note that the system still needs to make a choice between senses in this setting since all senses are represented in the model, including senses that do not occur in the test data. We also see that defining a lower bound for senses not occurring in the test set makes a difference. If we define the lower bound to 5, performance drops by 4 to 5 points, confirming the earlier experiment in which we just boost the LFS cases.

5 Discussion

In general, we note that adding more training data, obtained through manual annotation or unsupervised learning, will lead to results that exceed the standard system trained on SemCor, but will not transcend the performance to a level where we would consider the task solved. Interestingly, balancing the distribution to the task turns out to be very effective to boost the accuracy to an upper bound of around 85%, which is close to what other NLP tasks, such as entity detection and linking, achieve. In addition, the amount of training data to achieve this can be kept relatively low and can be acquired automatically. The main conclusion we draw from this is that test sets appear to contain very specific idiosyncratic details (long-tail details) when it comes to semantic tasks (as opposed to for instance syntactic tasks). These details are difficult to capture using the available training data in general. Just providing more does not help.

What helps is determining the semantic specifics of the test data. We believe that this is not just a matter of finding the predominant sense as argued by McCarthy et al. (2004a), since the distribution of both predominant and nondominant senses play a role. We have seen that also MFS distributions continue to play a role as they show shifts that need to be captured as well. We thus expect that acquisition, and likewise modeling, should be considered as problem-driven tasks rather than applying bulk acquisition. This is exactly what Figure 2 indicates. The improvements are for very specific low frequent senses and not for others. We believe that each test set has a unique long tail profile that needs to be captured by the system. In future work, we hope to perform well on this task without any prior knowledge on the sense distribution of the task, but by analyzing the properties of the texts.

The results from our experiment show that a lot of recent silver data results in better performance than a small amount of old manually-annotated data. It would be interesting to perform experiments with old silver data and manually annotated recent data. The practical obstacle towards achieving this is the availability of the data. For manual annotation, this would require a big annotation effort. For silver data, this would require an innovative approach that is different from recent approaches, which are mostly based on Wikipedia, e.g. Babelfied Wikipedia (Scozzafava et al., 2015).

One could argue that our system just uses the prior probabilities and that the machine learning is not adding anything on top of this. This is undeniable a factor but we also see that the Provenance of the training data has some impact, as the WW data help more than the GC data. Furthermore, it is not the case that the lemmas in the test set only occur in a single sense. The system does need to make a choice between different instances of a lemma in the test set for 41.4% of the lemmas.

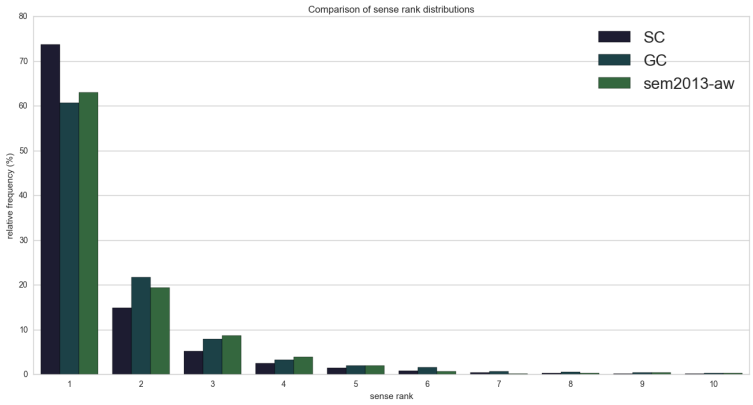

Figure 2: Comparison of sense rank distributions from the corpora: SC (left column), GC (middle column), and sem2013-aw (right column) for the sense ranks 1 till 10.

To our knowledge, only two systems (out-of-competition) achieved a better accuracy score on the **sem2013-aw** competition compared to our best run, experiment 4, from the unbalanced systems (experiments 1 to 7). The Game-Theoretic approach to WSD in Tripodi and Pelillo (2016) leads to an accuracy of 70.8, which is achieved by not only relying on the sentence as the context, but on the full document. Weissenborn et al. (2015) achieve the best result to date with an accuracy of 72.28%. The improvement is mainly due to the joint disambiguation of nouns and entities. Besides the fact that our system is more basic and uses a poorer context, we also see that even a simple system as ours can achieve far better results if we would know the distribution of the test set. We can expect that more advanced systems such as Tripodi and Pelillo (2016) and Weissenborn et al. (2015) may even exceed our best results of 86.8 with perfect balancing.

For future work, we hence intend to develop models that are problem-driven and attempt to obtain meta properties of the target data to estimate better sense distributions. An interesting starting point is a Word Sense Induction system called HCA-WSI (Bennett et al., 2016). This system could be used to determine the sense distributions of the time period in which a document in the test data was written, which would remove the dependence on sense distributions from manually annotated corpora.

6 Conclusions

We addressed the problem that most WSD systems perform well on the MFS and extremely poorly on the LFS, due to the skewness of the training data to the MFS. We analyzed the impact of adapting the training data with regard to this skewed performance: more data, better data, and balanced data. In general, we observe that more training data does indeed improve the results. However, the provenance of the data proved to be more important than the volume. We observed that automatically-acquired training data that was closer to the test data with respect to time and genre yielded better results than manually-created training data from the WordNet glosses. Finally, the real improvement was found by mimicking the sense distribution of the test data in the training data, which lead to results close to where we would consider the task solved. Hence, we argue that the solution to the task can be found in problem-driven approaches that attempt to adapt their models with respect to properties of the test set. In future work, we will focus on unsupervised methods to identify the meta properties of a test set and to capture its idiosyncratic long-tail details.

References

Eneko Agirre and Oier Lopez de Lacalle. 2008. On robustness and domain adaptation using svd for word sense disambiguation. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 17–24. Coling 2008 Organizing Committee.

Eneko Agirre and David Martinez, 2004. *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, chapter Unsupervised WSD based on Automatically Retrieved Examples: The Importance of Bias.

Eneko Agirre, Oier López de Lacalle, Christiane Fellbaum, Shu-Kai Hsieh, Maurizio Tesconi, Monica Monachini, Piek Vossen, and Roxanne Segers. 2010. SemEval-2010 Task 17: All-Words Word Sense Disambiguation on a Specific Domain. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 75–80. Association for Computational Linguistics.

Andrew Bennett, Timothy Baldwin, Han Jey Lau, Diana McCarthy, and Francis Bond. 2016. Lexsemtm: A semantic dataset based on all-words unsupervised sense distribution learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1513–1524. Association for Computational Linguistics.

David Blei, Andrew Ng, and Michael Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Jordan Boyd-Graber and David Blei. 2007. Putop: Turning predominant senses into a topic model for word sense disambiguation. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 277–281. Association for Computational Linguistics.

Hiram Calvo and Alexander Gelbukh. 2015. Is the Most Frequent Sense of a Word Better Connected in a Semantic Network? In *Advanced Intelligent Computing Theories and Applications*, pages 491–499. Springer.

Marine Carpuat, Hal Daume III, Katharine Henry, Ann Irvine, Jagadeesh Jagarlamudi, and Rachel Rudinger. 2013. Sensespotting: Never let your parallel data tie you to an old domain. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1435–1445. Association for Computational Linguistics.

Yee Seng Chan and Hwee Tou Ng. 2005. Word Sense Disambiguation with Distribution Estimation. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, IJCAI'05, pages 1010–1015, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Hal Daume III. 2007. Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL 2007)*, pages 256–263, Prague, Czech Republic. Association for Computational Linguistics.

Christiane Fellbaum, editor. 1998. *WordNet An Electronic Lexical Database*. The MIT Press, Cambridge, MA ; London, May.

Winthrop Nelson Francis and Henry Kucera. 1979. Brown corpus manual. *Brown University*.

Yoan Gutiérrez, Yenier Castañeda, Andy González, Rainel Estrada, D. Dennys Piug, I. Jose Abreu, Roger Pérez, Antonio Fernández Orquín, Andrés Montoyo, Rafael Muñoz, and Franc Camara. 2013. Umcc_dlsi: Reinforcing a ranking algorithm with sense frequencies and multidimensional semantic resources to solve multilingual word sense disambiguation. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 241–249. Association for Computational Linguistics.

Jing Jiang and ChengXiang Zhai. 2007. Instance weighting for domain adaptation in nlp. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 264–271. Association for Computational Linguistics.

Adam Kilgarriff, 2004. *How Dominant Is the Commonest Sense of a Word?*, pages 103–111. Springer Berlin Heidelberg, Berlin, Heidelberg.

Rob Koeling, Diana McCarthy, and John Carroll. 2005. Domain-specific sense distributions and predominant sense acquisition. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*.

Kazuhei Kouno, Hiroyuki Shinnou, Minoru Sasaki, and Kanako Komiya. 2015. Unsupervised domain adaptation for word sense disambiguation using stacked denoising autoencoder. In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation: Posters*, pages 224–231.

Han Jey Lau, Paul Cook, Diana McCarthy, David Newman, and Timothy Baldwin. 2012. Word sense induction for novel sense detection. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 591–601. Association for Computational Linguistics.

Han Jey Lau, Paul Cook, Diana McCarthy, Spandana Gella, and Timothy Baldwin. 2014. Learning word sense distributions, detecting unattested senses and identifying novel senses using topic models. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 259–270. Association for Computational Linguistics.

Diana McCarthy, Rob Koeling, Julie Weeds, and John Carroll. 2004a. Finding predominant word senses in untagged text. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*.

Diana McCarthy, Rob Koeling, Julie Weeds, and John Carroll. 2004b. Proceedings of senseval-3, the third international workshop on the evaluation of systems for the semantic analysis of text.

Diana McCarthy, Rob Koeling, Julie Weeds, and John Carroll. 2007. Unsupervised acquisition of predominant word senses. *Computational Linguistics, Volume 33, Number 4, December 2007*.

George Miller, Claudia Leacock, Randee Tengi, and Ross T. Bunker. 1993. A semantic concordance. In *HUMAN LANGUAGE TECHNOLOGY: Proceedings of a Workshop Held at Plainsboro, New Jersey, March 21-24, 1993*.

Andrea Moro, Francesco Cecconi, and Roberto Navigli. 2014. Multilingual word sense disambiguation and entity linking for everybody. In *Proceedings of the 2014 International Conference on Posters & Demonstrations Track-Volume 1272*, pages 25–28. CEUR-WS.org.

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.

Roberto Navigli, David Jurgens, and Daniele Vannella. 2013. SemEval-2013 task 12: Multilingual Word Sense Disambiguation. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 222–231, Atlanta, Georgia, USA, June. Association for Computational Linguistics.

Mark EJ Newman. 2005. Power laws, Pareto distributions and Zipf's law. *Contemporary physics*, 46(5):323–351.

Marten Postma, Ruben Izquierdo, Eneko Agirre, German Rigau, and Piek Vossen. 2016. Addressing the MFS Bias in WSD systems. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).

Judita Preiss. 2006. A detailed comparison of WSD systems: an analysis of the system answers for the Senseval-2 English all words task. *Natural Language Engineering*, 12(03):209–228.

Federico Scozzafava, Alessandro Raganato, Andrea Moro, and Roberto Navigli. 2015. Automatic Identification and Disambiguation of Concepts and Named Entities in the Multilingual Wikipedia. In *Congress of the Italian Association for Artificial Intelligence*, pages 357–366. Springer.

Yee Whye Teh, Michael Jordan, Matthew Beal, and David Blei. 2012. Hierarchical dirichlet processes. *Journal of the American Statistical Association*.

Rocco Tripodi and Marcello Pelillo. 2016. A game-theoretic approach to word sense disambiguation. *arXiv preprint arXiv:1606.07711*.

Dirk Weissenborn, Leonhard Hennig, Feiyu Xu, and Hans Uszkoreit. 2015. Multi-objective optimization for the joint disambiguation of nouns and named entities. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 596–605. Association for Computational Linguistics.

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