Relieving The Data Acquisition Bottleneck In Word Sense Disambiguation

Mona Diab
Linguistics Department
Stanford University
mdiab@stanford.edu

Abstract

Supervised learning methods for WSD yield better performance than unsupervised methods. Yet the availability of clean training data for the former is still a severe challenge. In this paper, we present an unsupervised bootstrapping approach for WSD which exploits huge amounts of automatically generated noisy data for training within a supervised learning framework. The method is evaluated using the 29 nouns in the English Lexical Sample task of SENSEVAL2. Our algorithm does as well as supervised algorithms on 31% of this test set, which is an improvement of 11% (absolute) over state-of-the-art bootstrapping WSD algorithms. We identify seven different factors that impact the performance of our system.

1 Introduction

Supervised Word Sense Disambiguation (WSD) systems perform better than unsupervised systems. But lack of training data is a severe bottleneck for supervised systems due to the extensive labor and cost involved. Indeed, one of the main goals of the SENSEVAL exercises is to create large amounts of sense-annotated data for supervised systems (Kilgarriff&Rosenzweig, 2000). The problem is even more challenging for languages which possess scarce computer readable knowledge resources.

In this paper, we investigate the role of large amounts of noisily sense annotated data obtained using an unsupervised approach in relieving the data acquisition bottleneck for the WSD task. We bootstrap a supervised learning WSD system with an unsupervised seed set. We use the sense annotated data produced by Diab's unsupervised system SALAAM (Diab&Resnik, 2002; Diab, 2003). SALAAM is a WSD system that exploits parallel corpora for sense disambiguation of words in running text. To date, SALAAM yields the best scores for an unsupervised system on the SENSEVAL2 English All-Words task (Diab, 2003). SALAAM is an appealing approach as it provides automatically sense annotated data in two languages simultaneously, thereby providing a multilingual framework for solving the data acquisition problem. For instance, SALAAM has been used to bootstrap the WSD process for Arabic as illustrated in (Diab, 2004).

In a supervised learning setting, WSD is cast as a classification problem, where a predefined set of sense tags constitutes the classes. The ambiguous words in text are assigned one or more of these classes by a machine learning algorithm based on some extracted features. This algorithm learns parameters from explicit associations between the class and the features, or combination of features, that characterize it. Therefore, such systems are very sensitive to the training data, and those data are, generally, assumed to be as clean as possible.

In this paper, we question that assumption. Can large amounts of noisily annotated data used in training be useful within such a learning paradigm for WSD? What is the nature of the quality-quantity trade-off in addressing this problem?

2 Related Work

To our knowledge, the earliest study of bootstrapping a WSD system with noisy data is by Gale et. al., (Gale et al. , 1992). Their investigation was limited in scale to six data items with two senses each and a bounded number of examples per test item.

Two more recent investigations are by Yarowsky, (Yarowsky, 1995), and later, Mihalcea, (Mihalcea, 2002). Each of the studies, in turn, addresses the issue of data quantity while maintaining good quality training examples. Both investigations present algorithms for bootstrapping supervised WSD systems using clean data based on a dictionary or an ontological resource. The general idea is to start with a clean initial seed and iteratively increase the seed size to cover more data.

Yarowsky starts with a few tagged instances to train a decision list approach. The initial seed is manually tagged with the correct senses based on entries in Roget's Thesaurus. The approach

yields very successful results — 95% — on a handful of data items.

Mihalcea, on the other hand, bases the bootstrapping approach on a generation algorithm, GenCor (Mihalcea&Moldovan, 1999). GenCor creates seeds from monosemous words in WordNet, Semcor data, sense tagged examples from the glosses of polysemous words in WordNet, and other hand tagged data if available. This initial seed set is used for querying the Web for more examples and the retrieved contexts are added to the seed corpus. The words in the contexts of the seed words retrieved are then disambiguated. The disambiguated contexts are then used for querying the Web for yet more examples, and so on. It is an iterative algorithm that incrementally generates large amounts of sense tagged data. The words found are restricted to either part of noun compounds or internal arguments of verbs. Mihalcea's supervised learning system is an instance-based-learning algorithm. In the study, Mihalcea compares results yielded by the supervised learning system trained on the automatically generated data, GenCor, against the same system trained on manually annotated data. She reports successful results on six of the data items tested.

3 Empirical Layout

Similar to Mihalcea's approach, we compare results obtained by a supervised WSD system for English using manually sense annotated training examples against results obtained by the same WSD system trained on SALAAM sense tagged examples. The test data is the same, namely, the SENSEVAL 2 English Lexical Sample test set. The supervised WSD system chosen here is the University of Maryland System for SENSEVAL 2 Tagging ($UMSST$) (Cabezas et al. , 2002).

3.1 $UMSST$

The learning approach adopted by $UMSST$ is based on Support Vector Machines (SVM). $UMSST$ uses *SVM-light*^{TM} by Joachims (Joachims, 1998).[1]

For each *target* word, where a target word is a test item, a family of classifiers is constructed, one for each of the target word senses. All the positive examples for a sense (S_i) are considered the negative examples of (S_j), where $i \neq j$.(Allwein et al., 2000) In $UMSST$, each target word is considered an independent classification problem.

The features used for $UMSST$ are mainly contextual features with weight values associated with each feature. The features are space delimited units, tokens, extracted from the immediate context of the target word. Three types of features are extracted:

- Wide Context Features: All the tokens in the paragraph where the target word occurs.
- Narrow Context features: The tokens that collocate in the surrounding context, to the left and right, with the target word within a fixed window size of 3.
- Grammatical Features: Syntactic tuples such as *verb-obj, subj-verb, etc.* extracted from the context of the target word using a dependency parser, MINIPAR (Lin, 1998).

Each feature extracted is associated with a weight value. The weight calculation is a variant on the *Inverse Document Frequency (IDF)* measure in Information Retrieval. The weighting, in this case, is an *Inverse Category Frequency (ICF)* measure where each token is weighted by the inverse of its frequency of occurrence in the specified context of the target word.

3.1.1 Manually Annotated Training Data

The manually-annotated training data is the SENSEVAL2 Lexical Sample training data for the English task, (SV2LS_Train).[2] This training data corpus comprises 44856 lines and 917740 tokens.

There is a close affinity between the test data and the manually annotated training data. The Pearson (r) correlation between the sense distributions for the test data and the manually annotated training data, per test item, ranges between $0.9 \leq 1$.[3]

3.2 SALAAM

SALAAM exploits parallel corpora for sense annotation. The key intuition behind SALAAM is that when words in one language, L1, are translated into the same word in a second language, L2, then those L1 words are semantically similar. For example, when the English — L1 — words *bank, brokerage, mortgage-lender* translate into the French — L2 — word *banque* in a parallel corpus, where *bank* is polysemous, SALAAM discovers that the intended sense for *bank* is the *financial institution* sense, not the *geological formation* sense, based on the fact that it is grouped with *brokerage* and *mortgage-lender*. SALAAM's algorithm is as follows:

- SALAAM expects a word aligned parallel corpus as input;

[1] http://www.ai.cs.uni.dortmund.de/svmlight.

[2] http://www.senseval.org

[3] The correlation is measured between two frequency distributions. Throughout this paper, we opt for using the parametric Pearson r correlation rather than KL distance in order to test statistical signifi cance.

- L1 words that translate into the same L2 word are grouped into clusters;
- SALAAM identifies the appropriate senses for the words in those clusters based on the words senses' proximity in WordNet. The word sense proximity is measured in information theoretic terms based on an algorithm by Resnik (Resnik, 1999);
- A sense selection criterion is applied to choose the appropriate sense label or set of sense labels for each word in the cluster;
- The chosen sense tags for the words in the cluster are propagated back to their respective contexts in the parallel text. Simultaneously, SALAAM projects the propagated sense tags for L1 words onto their L2 corresponding translations.

3.2.1 Automatically Generated SALAAM Training Data

Three sets of SALAAM tagged training corpora are created:

- **SV2LS_TR**: English SENSEVAL2 Lexical Sample trial and training corpora with no manual annotations. It comprises 61879 lines and 1084064 tokens.
- **MT**: The English Brown Corpus, SENSEVAL1 (trial, training and test corpora), Wall Street Journal corpus, and SENSEVAL 2 All Words corpus. All of which comprise 151762 lines and 37945517 tokens.
- **HT**: UN English corpus which comprises 71672 lines of 1734001 tokens

The SALAAM-tagged corpora are rendered in a format similar to that of the manually annotated training data. The automatic sense tagging for MT and SV2LS_TR training data is based on using SALAAM with machine translated parallel corpora. The HT training corpus is automatically sense tagged based on using SALAAM with the English-Spanish UN naturally occurring parallel corpus.

3.3 Experimental Conditions

Experimental conditions are created based on three of SALAAM's tagging factors, **Corpus, Language** and **Threshold**:

- **Corpus**: There are 4 different combinations for the training corpora: MT+SV2LS_TR; MT+HT+SV2LS_TR; HT+SV2LS_TR; or SV2LS_TR alone.
- **Language**: The context language of the parallel corpus used by SALAAM to obtain the sense tags for the English training corpus. There are three options: French (FR), Spanish (SP), or, Merged languages (ML), where the results are obtained by merging the English output of FR and SP.
- **Threshold**: Sense selection criterion, in SALAAM, is set to either MAX (M) or THRESH (T).

These factors result in 39 conditions.[4]

3.4 Test Data

The test data are the 29 noun test items for the SENSEVAL 2 English Lexical Sample task, (SV2LS-Test). The data is tagged with the WordNet 1.7pre (Fellbaum, 1998; Cotton et al. , 2001). The average perplexity for the test items is 3.47 (see Section 5.3), the average number of senses is 7.93, and the total number of contexts for all senses of all test items is 1773.

4 Evaluation

In this evaluation, $UMSST\_S$ is the $UMSST$ system trained with SALAAM-tagged data and $UMSST\_H$ is the $UMSST$ system trained with manually annotated data. Since we don't expect $UMSST\_S$ to outperform human tagging, the results yielded by $UMSST\_H$, are the upper bound for the purposes of this study. It is important to note that $UMSST\_S$ is always trained with **SV2LS_TR** as part of the training set in order to guarantee genre congruence between the training and test sets.The scores are calculated using `scorer2`.[5] The average precision score over all the items for $UMSST\_H$ is 65.3% at 100% Coverage.

4.1 Metrics

We report the results using two metrics, the harmonic mean of precision and recall, ($F_{\beta=1}$) score, and the Performance Ratio (PR), which we define as the ratio between two precision scores on the same test data where precision is rendered using `scorer2`. PR is measured as follows:

$$PR = \frac{\texttt{UMSST\_S}Precision}{\texttt{UMSST\_H}Precision} \quad (1)$$

[4] Originally, there are 48 conditions, 9 of which are excluded due to extreme sparseness in training contexts.

[5] From http://www.senseval.org, all `scorer2` results are reported in fi ne-grain mode.

4.2 Results

Table 1 shows the $F_{\beta=1}$ scores for the upper bound $UMSST\_H$. $UMSST\_S_{best}$ is the condition in $UMSST\_S$ that yields the highest overall $F_{\beta=1}$ score over all noun items. $UMSST\_S_{max}$ the maximum $F_{\beta=1}$ score achievable, if we know which condition yields the best performance per test item, therefore it is an oracle condition.[6] Since our approach is unsupervised, we also report the results of other unsupervised systems on this test set. Accordingly, the last seven row entries in Table 1 present state-of-the-art SENSEVAL2 *unsupervised* systems performance on this test set.[7]

System	$F_{\beta=1}$
$UMSST\_H$	65.3
$UMSST\_S_{best}$	36.02
$UMSST\_S_{max}$	45.1
ITRI	45
UNED-LS-U	40.1
CLRes	29.3
IIT2(R)	24.4
IIT1(R)	23.9
IIT2	23.2
IIT1	22

Table 1: $F_{\beta=1}$ scores on SV2LS_Test for $UMSST\_H$, $UMSST\_S_{best}$, $UMSST\_S_{max}$, and state-of-the-art unsupervised systems participating in the SENSEVAL2 English Lexical Sample task.

All of the unsupervised methods including $UMSST\_S_{best}$ and $UMSST\_S_{max}$ are significantly below the supervised method, $UMSST\_H$. $UMSST\_S_{best}$ is the third in the unsupervised methods. It is worth noting that the average $F_{\beta=1}$ score across the 39 conditions is 33.64, and the lowest is 31.16. The five best conditions for $UMSST\_S$, that yield the highest average $F_{\beta=1}$ across all test items, use the HT corpus in the training data, four of which are the result of merged languages in SALAAM indicating that evidence from different languages simultaneously is desirable. $UMSST\_S_{max}$ is the maximum potential among all unsupervised approaches if the best of all the conditions are combined. One of our goals is to automatically determine which condition or set of conditions yield the best results for each test item.

Of central interest in this paper is the performance ratio (PR) for the individual nouns. Table 2 illustrates the PR of the different nouns yielded by $UMSST\_S_{best}$ and $UMSST\_S_{max}$ sorted in descending order by $UMSST\_S_{max}$ PR scores. A 1.00 PR indicates an equivalent performance between $UMSST\_S$ and $UMSST\_H$. The highest PR values are highlighted in bold.

Nouns	#Ss	UMH%	UMSb	UMSm
detention	4	65.6	**1.00**	**1.05**
chair	7	83.3	**1.02**	**1.02**
bum	4	85	0.14	**1.00**
dyke	2	89.3	**1.00**	**1.00**
fatigue	6	80.5	**1.00**	**1.00**
hearth	3	75	**1.00**	**1.00**
spade	6	75	**1.00**	**1.00**
stress	6	50	0.05	**1.00**
yew	3	78.6	**1.00**	**1.00**
art	17	47.9	**0.98**	**0.98**
child	7	58.7	**0.93**	**0.97**
material	16	55.9	0.81	**0.92**
church	6	73.4	0.75	0.77
mouth	10	55.9	0	0.73
authority	9	62	0.60	0.70
post	12	57.6	0.66	0.66
nation	4	78.4	0.34	0.59
feeling	5	56.9	0.33	0.59
restraint	8	60	0.2	0.56
channel	7	62	0.52	0.52
facility	5	54.4	0.32	0.51
circuit	13	62.7	0.44	0.44
nature	7	45.7	0.43	0.43
bar	19	60.9	0.20	0.30
grip	6	58.8	0.27	0.27
sense	8	39.6	0.24	0.24
lady	8	72.7	0.09	0.16
day	16	62.5	0.06	0.08
holiday	6	86.7	0.08	0.08

Table 2: The number of senses per item, in column #Ss, $UMSST\_H$ precision performance per item as indicated in column UMH, PR scores for $UMSST\_S_{best}$ in column UMSb and $UMSST\_S_{max}$ in column UMSm on SV2LS_Test

$UMSST\_S_{max}$ yields PR scores > 0.91 for the top 12 test items listed in Table 2. Our algorithm does as well as supervised algorithm, $UMSST\_H$, on 41.6% of this test set. In $UMSST\_S_{best}$, 31% of the test items, (9 nouns yield PR scores > 0.92), do as well as $UMSST\_H$. This is an improvement of 11% absolute over state-of-the-art bootstrapping WSD algorithm yielded by Mihalcea (Mihalcea, 2002). Mihalcea reports high PR scores for six test items only: *art, chair, channel, church, detention, nation*. It is worth highlighting that her bootstrapping approach is partially supervised since

[6] The different conditions are considered independent taggers and there is no interaction across target nouns

[7] http://www.senseval.org

it depends mainly on hand labelled data as a seed for the training data.

Interestingly, two nouns, *detention* and *chair*, yield better performance than $UMSST\_H$, as indicated by the PRs 1.05 and 1.02, respectively. This is attributed to the fact that SALAAM produces a lot more correctly annotated training data for these two words than that provided in the manually annotated training data for $UMSST\_H$.

Some nouns yield very poor PR values mainly due to the lack of training contexts, which is the case for *mouth* in $UMSST\_S_{best}$, for example. Or lack of coverage of all the senses in the test data such as for *bar* and *day*, or simply errors in the annotation of the SALAAM-tagged training data.

If we were to include only nouns that achieve acceptable PR scores of ≥ 0.65 — the first 16 nouns in Table 2 for $UMSST\_S_{max}$ — the overall potential precision of $UMSST\_S$ is significantly increased to 63.8% and the overall precision of $UMSST\_H$ is increased to 68.4%.[8]

These results support the idea that we could replace hand tagging with SALAAM's unsupervised tagging if we did so for those items that yield an acceptable PR score. But the question remains: How do we predict which training/test items will yield acceptable PR scores?

5 Factors Affecting Performance Ratio

In an attempt to address this question, we analyze several different factors for their impact on the performance of $UMSST\_S$ quanitified as PR. In order to effectively alleviate the sense annotation acquisition bottleneck, it is crucial to predict which items would be reliably annotated automatically using $UMSST\_S$. Accordingly, in the rest of this paper, we explore 7 different factors by examining the yielded PR values in $UMSST\_S_{max}$.

5.1 Number of Senses

The test items that possess many senses, such as *art* (17 senses), *material* (16 senses), *mouth* (10 senses) and *post* (12 senses), exhibit PRs of 0.98, 0.92, 0.73 and 0.66, respectively. Overall, the correlation between number of senses per noun and its PR score is an insignificant $r = -0.31$, $(F(1, 27) = 2.9, p > 0.1)$. Though it is a weak negative correlation, it does suggest that when the number of senses increases, PR tends to decrease.

5.2 Number of Training Examples

This is a characteristic of the training data. We examine the correlation between the PR and the number of training examples available to $UMSST\_S$ for each noun in the training data. The correlation between the number of training examples and PR is insignificant at $r = -0.15$, $(F(1, 27) = 0.637, p > 0.4)$. More interestingly, however, *spade*, with only 5 training examples, yields a PR score of 1.0. This contrasts with *nation*, which has more than 4200 training examples, but yields a low PR score of 0.59. Accordingly, the number of training examples alone does not seem to have a direct impact on PR.

5.3 Sense Perplexity

This factor is a characteristic of the training data. Perplexity is $2^{Entropy}$. Entropy is measured as follows:

$$H(X) = -\sum_{x \in X} p(x) log_2(p(x)) \qquad (2)$$

where x is a sense for a polysemous noun and X is the set of all its senses.

Entropy is a measure of confusability in the senses' contexts distributions; when the distribution is relatively uniform, entropy is high. A skew in the senses' contexts distributions indicates low entropy, and accordingly, low perplexity. The lowest possible perplexity is 1, corresponding to 0 entropy. A low sense perplexity is desirable since it facilitates the discrimination of senses by the learner, therefore leading to better classification. In the SALAAM-tagged training data, for example, *bar* has the highest perplexity value of 9.85 over its 19 senses, while *day*, with 16 senses, has a much lower perplexity of 1.3.

Surprisingly, we observe nouns with high perplexity such as *bum* (sense perplexity value of 3.03) achieving PR scores of 1.0. While nouns with relatively low perplexity values such as *grip* (sense perplexity of 0.53) yields a low PR score of 0.26. Moreover, nouns with the same perplexity and similar number of senses yield very different PR scores. For example, examining *holiday* and *child*, both have the same perplexity of 2.144 and the number of senses is close, with 6 and 7 senses, respectively, however, the PR scores are very different; *holiday* yields a PR of 0.08, and *child* achieves a PR of 0.97. Furthermore, *nature* and *art* have the same perplexity of 2.29; *art* has 17 senses while *nature* has 7 senses only, nonetheless, *art* yields a much higher PR score of (0.98) compared to a PR of 0.44 for *nature*.

These observations are further solidified by the insignificant correlation of $r = -0.12$, $(F(1, 27) = 0.45, p > 0.5)$ between sense perplexity and PR.

At first blush, one is inclined to hypothesize that,

[8] A PR of 0.65 is considered acceptable since $UMSST\_H$ achieves an overall $F_{\beta=1}$ score of 65.3 in the WSD task.

the combination of low perplexity associated with a large number of senses — as an indication of high skew in the distribution — is a good indicator of high PR, but reviewing the data, this hypothesis is dispelled by *day* which has 16 senses and a sense perplexity of 1.3, yet yields a low PR score of 0.08.

5.4 Semantic Translation Entropy

Semantic translation entropy (STE) (Melamed, 1997) is a special characteristic of the SALAAM-tagged training data, since the source of evidence for SALAAM tagging is multilingual translations. STE measures the amount of translational variation for an L1 word in L2, in a parallel corpus. STE is a variant on the entropy measure. STE is expressed as follows:

$$H(T|s) = -\sum_{t \in T} p(t|s).log_2(p(t|s)) \qquad (3)$$

where t is a translation in the set of possible translations T in L2; and s is L1 word.

The probability of a translation t is calculated directly from the alignments of the test nouns and their corresponding translations via the maximum likelihood estimate.

Variation in translation is beneficial for SALAAM tagging, therefore, high STE is a desirable feature. Correlation between the automatic tagging precision and STE is expected to be high if SALAAM has good quality translations and good quality alignments. However, this correlation is a low $r = 0.33$. Consequently, we observe a low correlation between STE and PR, $r = 0.22$, $(F(1, 27) = 1.31, p > 0.26)$.

Examining the data, the nouns *bum, detention, dyke, stress,* and *yew* exhibit both high STE and high PR; Moreover, there are several nouns that exhibit low STE and low PR. But the intriguing items are those that are inconsistent. For instance, *child* and *holiday*: *child* has an STE of 0.08 and comprises 7 senses at a low sense perplexity of 1.69, yet yields a high PR of 0.97. As mentioned earlier, low STE indicates lack of translational variation. In this specific experimental condition, *child* is translated as {*enfant, enfantile, niño, niño-pequeño*}, which are words that preserve ambiguity in both French and Spanish. On the other hand, *holiday* has a relatively high STE value of 0.66, yet results in the lowest PR of 0.08. Consequently, we conclude that STE alone is not a good direct indicator of PR.

5.5 Perplexity Difference

Perplexity difference (PerpDiff) is a measure of the absolute difference in sense perplexity between the test data items and the training data items. For the manually annotated training data items, the overall correlation between the perplexity measures is a significant $r = 0.96$ which contrasts to a low overall correlation of $r = 0.43$ between the SALAAM-tagged training data items and the test data items. Across the nouns in this study, the correlation between PerpDiff and PR is $r = -0.4$. It is advantageous to be as similar as possible to the training data to guarantee good classification results within a supervised framework, therefore a low PerpDiff is desirable. We observe cases with a low PerpDiff such as *holiday* (PerpDiff of 0.05), yet the PR is a low 0.08. On the other hand, items such as *art* have a relatively high PerpDiff of 2.62, but achieves a high PR of 0.97. Accordingly, PerpDiff alone is not a good indicator of PR.

5.6 Sense Distributional Correlation

Sense Distributional Correlation (SDC) results from comparing the sense distributions of the test data items with those of SALAAM-tagged training data items. It is worth noting that the correlation between the SDC of manually annotated training data and that of the test data ranges from $r = 0.9 - 1.0$. A strong significant correlation of $r = 0.87$, $(F(1, 27) = 80, p < 0.0001)$ between SDC and PR exists for SALAAM-tagged training data and the test data. Overall, nouns that yield high PR have high SDC values. However, there are some instances where this strong correlation is not exhibited. For example, *circuit* and *post* have relatively high SDC values, 0.794 and 0.859, respectively, in $UMSST\_S_{max}$, but they score lower PR values than *detention* which has a comparatively lower SDC value of 0.776. The fact that both *circuit* and *post* have many senses, 13 and 12, respectively, while *detention* has 4 senses only is noteworthy. *detention* has a higher STE and lower sense perplexity than either of them however. Overall, the data suggests that SDC is a very good direct indicator of PR.

5.7 Sense Context Confusability

A situation of sense context confusability (SCC) arises when two senses of a noun are very similar and are highly uniformly represented in the training examples. This is an artifact of the fine granularity of senses in WordNet 1.7pre. Highly similar senses typically lead to similar usages, therefore similar contexts, which in a learning framework detract from the learning algorithm's discriminatory power.

Upon examining the 29 polysemous nouns in the training and test sets, we observe that a significant number of the words have similar senses according

to a manual grouping provided by Palmer, in 2002.[9] For example, senses 2 and 3 of *nature*, meaning *trait* and *quality*, respectively, are considered similar by the manual grouping. The manual grouping does not provide total coverage of all the noun senses in this test set. For instance, it only considers the homonymic senses 1, 2 and 3 of *spade*, yet, in the current test set, *spade* has 6 senses, due to the existence of sub senses.

26 of the 29 test items exhibit multiple groupings based on the manual grouping. Only three nouns, *detention, dyke, spade* do not have any sense groupings. They all, in turn, achieve high PR scores of 1.0.

There are several nouns that have relatively high SDC values yet their performance ratios are low such as *post, nation, channel* and *circuit*. For instance, *nation* has a very high SDC value of 0.962, a low sense perplexity of 1.3 — relatively close to the 1.6 sense perplexity of the test data — a sufficient number of contexts (4350), yet it yields a PR of 0.59. According to the manual sense grouping, senses 1 and 3 are similar, and indeed, upon inspection of the context distributions, we find the bulk of the senses' instance examples in the SALAAM-tagged training data for the condition that yields this PR in $UMSST\_S_{max}$ are annotated with either sense 1 or sense 3, thereby creating confusable contexts for the learning algorithm. All the cases of nouns that achieve high PR and possess sense groups do not have any SCC in the training data which strongly suggests that SCC is an important factor to consider when predicting the PR of a system.

5.8 Discussion

We conclude from the above exploration that SDC and SCC affect PR scores directly. PerpDiff, STE, and Sense Perplexity, number of senses and number of contexts seem to have no noticeable direct impact on the PR.

Based on this observation, we calculate the SDC values for all the training data used in our experimental conditions for the 29 test items.

Table 3 illustrates the items with the highest SDC values, in descending order, as yielded from any of the SALAAM conditions. We use an empirical cut-off value of 0.75 for SDC. The SCC values are reported as a boolean Y/N value, where a Y indicates the presence of a sense confusable context. As shown a high SDC can serve as a means of automatically predicting a high PR, but it is not sufficient. If we eliminate the items where an SCC exists, namely, *mouth, post,* and *authority*, we are still left with *nation* and *circuit*, where both yield very low PR scores. *nation* has the desirable low PerpDiff of 0.22. The sense annotation tagging precision of the $SV2LS\_TR$ in this condition which yields the highest SDC — Spanish UN data with the $SV2LS\_TR$ for training — is a low 30.4% and a low STE value of 0.129. This is due to the fact that both French and Spanish preserve ambiguity in similar ways to English which does not make it a good target word for disambiguation within the SALAAM framework, given these two languages as sources of evidence. Accordingly, in this case, STE coupled with the noisy tagging could have resulted in the low PR. However, for *circuit*, the STE value for its respective condition is a high 0.291, but we observe a relatively high PerpDiff of 1.53 compared to the PerpDiff of 0 for the manually annotated data.

Noun	SDC	SCC	PR
dyke	1	N	1.00
bum	1	N	1.00
fatigue	1	N	1.00
hearth	1	N	1.00
yew	1	N	1.00
chair	0.99	N	1.02
child	0.99	N	0.95
detention	0.98	N	1.0
spade	0.97	N	1.00
mouth	0.96	Y	0.73
nation	0.96	N	0.59
material	0.92	N	0.92
post	0.90	Y	0.63
authority	0.86	Y	0.70
art	0.83	N	0.98
church	0.80	N	0.77
circuit	0.79	N	0.44
stress	0.77	N	1.00

Table 3: Highest SDC values for the test items associated with their respective SCC and PR values.[11]

Therefore, a combination of high SDC and nonexistent SCC can reliably predict good PR. But the other factors still have a role to play in order to achieve accurate prediction.

It is worth emphasizing that two of the identified factors are dependent on the test data in this study, SDC and PerpDiff. One solution to this problem is to estimate SDC and PerpDiff using a held out data set that is hand tagged. Such a held out data set would be considerably smaller than the required

[9] http://www.senseval.org/sense-groups. The manual sense grouping comprises 400 polysemous nouns including the 29 nouns in this evaluation.

size of a manually tagged training data for a classical supervised WSD system. Hence, SALAAM-tagged training data offers a viable solution to the annotation acquisition bottleneck.

6 Conclusion and Future Directions

In this paper, we applied an unsupervised approach within a learning framework $UMSST\_S$ for the sense annotation of large amounts of data. The ultimate goal of $UMSST\_S$ is to alleviate the data labelling bottleneck by means of a trade-off between quality and quantity of the training data. $UMSST\_S$ is competitive with state-of-the-art unsupervised systems evaluated on the same test set from SENSEVAL2. Moreover, it yields superior results to those obtained by the only comparable bootstrapping approach when tested on the same data set. Moreover, we explore, in depth, different factors that directly and indirectly affect the performance of $UMSST\_S$ quantified as a performance ratio, PR. Sense Distribution Correlation (SDC) and Sense Context Confusability (SCC) have the highest direct impact on performance ratio, PR. However, evidence suggests that probably a confluence of all the different factors leads to the best prediction of an acceptable PR value. An investigation into the feasibility of combining these different factors with the different attributes of the experimental conditions for SALAAM to automatically predict when the noisy training data can reliably replace manually annotated data is a matter of future work.

7 Acknowledgements

I would like to thank Philip Resnik for his guidance and insights that contributed tremendously to this paper. Also I would like to acknowledge Daniel Jurafsky and Kadri Hacioglu for their helpful comments. I would like to thank the three anonymous reviewers for their detailed reviews. This work has been supported, in part, by NSF Award #IIS-0325646.

References

Erin L. Allwein, Robert E. Schapire, and Yoram Singer. 2000. *Reducing multiclass to binary: A unifying approach for margin classifiers*. Journal of Machine Learning Research, 1:113-141.

Clara Cabezas, Philip Resnik, and Jessica Stevens. 2002. *Supervised Sense Tagging using Support Vector Machines*. Proceedings of the Second International Workshop on Evaluating Word Sense Disambiguation Systems (SENSEVAL-2). Toulouse, France.

Scott Cotton, Phil Edmonds, Adam Kilgarriff, and Martha Palmer, ed. 2001. *SENSEVAL-2: Second International Workshop on Evaluating Word Sense Disambiguation Systems*. ACL SIGLEX, Toulouse, France.

Mona Diab. 2004. *An Unsupervised Approach for Bootstrapping Arabic Word Sense Tagging*. Proceedings of Arabic Based Script Languages, COLING 2004. Geneva, Switzerland.

Mona Diab and Philip Resnik. 2002. *An Unsupervised Method for Word Sense Tagging Using Parallel Corpora*. Proceedings of 40th meeting of ACL. Pennsylvania, USA.

Mona Diab. 2003. *Word Sense Disambiguation Within a Multilingual Framework*. PhD Thesis. University of Maryland College Park, USA.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

William A. Gale and Kenneth W. Church and David Yarowsky. 1992. *Using Bilingual Materials to Develop Word Sense Disambiguation Methods*. Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation. Montréal, Canada.

Thorsten Joachims. 1998. *Text Categorization with Support Vector Machines: Learning with Many Relevant Features*. Proceedings of the European Conference on Machine Learning. Springer.

A. Kilgarriff and J. Rosenzweig. 2000. *Framework and Results for English SENSEVAL*. Journal of Computers and the Humanities. pages 15—48, 34.

Dekang Lin. 1998. *Dependency-Based Evaluation of MINIPAR*. Proceedings of the Workshop on the Evaluation of Parsing Systems, First International Conference on Language Resources and Evaluation. Granada, Spain.

Dan I. Melamed. 1997. *Measuring Semantic Entropy*. ACL SIGLEX, Washington, DC.

Rada Mihalcea and Dan Moldovan. 1999. *A method for Word Sense Disambiguation of unrestricted text*. Proceedings of the 37th Annual Meeting of ACL. Maryland, USA.

Rada Mihalcea. 2002. *Bootstrapping Large sense tagged corpora*. Proceedings of the 3rd International Conference on Languages Resources and Evaluations (LREC). Las Palmas, Canary Islands, Spain.

Philip Resnik. 1999. *Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language*. Journal Artifi cial Intelligence Research. (11) p. 95-130.

David Yarowsky. 1995. *Unsupervised Word Sense Disambiguation Rivaling Supervised Methods*. Proceedings of the 33rd Annual Meeting of ACL. Cambridge, MA.