Psycholinguistics, Lexicography, and Word Sense Disambiguation

Oi Yee Kwong
Department of Chinese, Translation and Linguistics
City University of Hong Kong
Tat Chee Avenue, Kowloon, Hong Kong
Olivia.Kwong@cityu.edu.hk

Abstract

Mainstream word sense disambiguation systems have relied mostly on supervised approaches. Complex interactions have been observed between learning algorithms and knowledge sources, but the factors underlying such phenomena are under-explored. This calls for more qualitative analysis of disambiguation results, possibly from an inter-disciplinary perspective. The current study thus preliminarily explores the relation between sense concreteness and the linguistic means for sense distinction with reference to the context availability model proposed in psycholinguistics and common practice in corpus-based lexicography. It will be shown that to a certain extent the varied usefulness of individual knowledge sources for target words, nouns in particular, may be related to the concreteness of the meanings concerned, which predicts how the sense is distinguished from other senses of the word in the first place. A better understanding of this relation is expected to inform the design of disambiguation systems which could then combine algorithms and knowledge sources in a genuine lexically sensitive way.

1 Introduction

Word sense ambiguities tend to escape people's awareness in everyday communication, except in deliberately biased artificial examples or when context is severely limited, since otherwise we almost effortlessly resolve them using a variety of linguistic and extra-linguistic knowledge. This wide range of information is often rendered as various knowledge sources in automatic word sense disambiguation (WSD) systems, partially modelled with different feature sets.

As exemplified in recent SENSEVAL and SEMEVAL evaluation exercises (e.g. Kilgarriff and Rosenzweig, 1999; Edmonds and Cotton, 2001; Mihalcea et al., 2004), state-of-the-art WSD systems are mostly based on supervised approaches. Machine learning algorithms are trained on sense-tagged examples, using a wide range of features extracted from the text approximating a variety of knowledge sources deemed useful for the purpose. Ensembles of different types of classifiers based on different feature sets with some voting scheme often report better performance than individual classifiers alone, though the advantage may just be marginal. While complex interactions between learning algorithms and knowledge sources have been observed (e.g. Mihalcea, 2002; Yarowsky and Florian, 2002), and although factors like sense granularity, availability of training data, part-of-speech (POS), etc. are found to relate to such interactions in one way or another, the nature underlying such interactions, which points to the lexical sensitivity issue of WSD, is still somehow under-explored. In particular, more qualitative analysis is needed for disambiguation results, possibly from an inter-disciplinary perspective, for a better understanding of the issue.

In the current study, we make a preliminary effort in this regard, and attempt to analyse disambiguation results with respect to the relation

26 th Pacific Asia Conference on Language, Information and Computation pages 408-417

between sense concreteness and the means for sense distinction in the first place. To this end, we refer to the context availability model proposed in psycholinguistics and common practice in modern corpus-based lexicography.

In Section 2, we will first briefly review related work with particular focus on the complex interaction between learning algorithms and knowledge sources in WSD revealed in recent evaluation exercises and various comparative studies, and present the Context Availability Model and discuss how it accounts for the concreteness effect in psycholinguistics. Section 3 reports on our qualitative analysis of the results from a simple WSD experiment on the noun samples in the SENSEVAL-3 English lexical sample task, for which we also made use of the Sketch Engine, a corpus query system popularly used in lexicography, as a tool for comparing the linguistic context availability among word senses. The paper will be concluded with future directions in Section 4.

2 Related Work

2.1 WSD: State of the Art

Two critical factors have been identified for the success of supervised WSD systems: learning algorithms and knowledge sources.

Individual learning algorithms are found to vary in their disambiguation performance. For instance, Màrquez et al. (2006) compared five machine learning algorithms widely used in previous studies, namely Naïve Bayes (NB), k-Nearest-Neighbor (kNN), Decision Lists (DL), AdaBoost (AB), and Support Vector Machines (SVM). They were trained on the same set of data and tested on examples selected from the DSO corpus. Knowledge sources were in the form of 15 local feature patterns (with words and POS) and topical context as bag of words (content words in the sentence). The most-frequent-sense classifier was used as baseline. It was found that all algorithms outperformed the baseline (46.55%), with SVM (67.07%) and AB performing significantly better than kNN, which in turn performed significantly better than NB and DL (61.34%).

Multiple knowledge sources are indispensable in WSD systems, and they contribute in different ways to disambiguation. Agirre and Stevenson (2006) summarised from many WSD studies the different knowledge sources available or extracted from various lexical resources and corpora, and their realisation as different features in individual systems. They generalised that all knowledge sources seem to provide useful disambiguation clues. Each POS profits from different knowledge sources, e.g. domain knowledge and topical word association are most useful for disambiguating nouns while local context benefits verbs and adjectives. The combination of all knowledge sources consistently gets the best results across POS categories. In addition, some learning algorithms are better suited to certain knowledge sources, and different grammatical categories may benefit from different learning algorithms.

Such a complex interaction between learning algorithms and knowledge sources was also exemplified in other comparative studies (e.g. Mihalcea, 2002; Yarowsky and Florian, 2002). The comprehensive study by Yarowsky and Florian (2002), for instance, compares the relative system performance across different training and data conditions with SENSEVAL-2 data on four languages. The results clearly show the interaction among feature sets, training sizes, and learning algorithms. They concluded that “there is no one-size-fits-all algorithm that excels at each of the diverse challenges in sense disambiguation”. For example, discriminative and aggregative algorithm classes often have complementary regions of effectiveness across numerous parameters, the former such as decision trees tend to perform well with local collocations or syntactic features, whereas the latter like Naïve Bayes tend to perform well with bag-of-word features. Some algorithms are more tolerant than others of sparse data, high degree of polysemy and noise in the training data.

2.2 The Lexical Sensitivity Issue

Despite such findings on the complex relationship between learning algorithms and knowledge sources, which possibly lead to the use of ensembles of classifiers with diverse knowledge sources in state-of-the-art systems, there are nevertheless some questions regarding their differential effectiveness left unanswered. One of the most important questions is how we could account for the intra-POS variation of the effectiveness of individual knowledge sources. Hence, while we find that target words of different

POS categories favour different knowledge sources for disambiguation, e.g. although local contexts are found to benefit verbs and adjectives more, they do contribute to the disambiguation of some nouns. What properties do such nouns possess? Can we predict the information susceptibility of individual words to optimize the use of different knowledge sources during disambiguation, and to consider the outcome given by different knowledge sources with different levels of confidence?

As Resnik and Yarowsky (1999) remarked, WSD is a highly lexically sensitive task which in effect requires specialized disambiguators for each polysemous word. But in what way precisely is the combination of algorithms and knowledge sources sensitive to individual (groups of) lexical items? Factors like the number of senses and how closely they are related will have an impact on the difficulty of disambiguation, and the varied difficulty may be reflected from the system performance (Chugur et al., 2002; Pedersen, 2002), but there is still more to learn, especially from an inter-disciplinary perspective. For instance, Krahmer (2010) encouraged mutual learning between computational linguists and psychologists, using as an example the possible influence of the general distinction between concrete and abstract language on perception shown in psychology studies, while such effects are somehow largely ignored in computational linguistics. We have also raised similar concerns for research on automatic word sense disambiguation (Kwong, 2012). In this study, we refer to the Context Availability Model in psycholinguistics (Schwanenflugel, 1991), which is used to explain human comprehension processes in general and more specifically to account for the concreteness effect in human word processing, to analyse WSD system performance on individual target words.

2.3 Context Availability Model

Polysemy, familiarity and concreteness have been considered important semantic characteristics which influence human lexical processing (e.g. Taft, 1991). While polysemy (in terms of sense number and granularity) and familiarity (in terms of frequency or prior probability) have also been addressed by computational linguists to account for differential system performance, the concreteness effect is somehow seldom discussed in the WSD literature. A few examples include: Jorgensen (1990) suggested that concreteness of a word may increase agreement between judges for sorting word usages and concrete words are easier to define; Kwong (2008) studied the relation between concreteness and system performance in SENSEVAL-2, though the findings were not particularly conclusive, partly because of the confusion from discussing concreteness at both the sense and word level; Yuret and Yatbaz (2010) mentioned that the abstract classes were responsible for most of the errors in their supersense tagging with unsupervised method. Given the significance of the concreteness effect in human lexical processing (e.g. Paivio et al., 1968; Kroll and Merves, 1986; Bleasdale, 1987; Schwanenflugel, 1991), more in-depth analysis of the concreteness effect is definitely needed especially for mainstream supervised WSD.

Psychologists have put forth various plausible explanations to account for the concreteness effect observed in human lexical processing, one of which is the context availability model. It suggests that the advantage of concrete words comes from their stronger and denser association to contextual knowledge than abstract words (Schwanenflugel 1991). The availability of contextual information enables a person to draw the relations between concepts that are needed for comprehension. Such contextual information may come from a person's prior knowledge or from the stimulus environment. According to this model, lexical decisions tend to take longer for abstract words because related contextual information that is used in deciding that an item is a word is less available for abstract words. Schwanenflugel et al. (1988) thus pointed out that the lexical decision times for abstract words are not necessarily longer than those for concrete words, especially when abstract concepts are also presented in relevant contexts. In addition, they found that rated context availability makes a better predictor for lexical decision time than imageability, familiarity, and age-of-acquisition. Thus the concreteness effects are rather attributable to the ease of retrieving related contextual information from prior knowledge for individual words, that is, context availability matters.

Such emphasis on the contextually based character of word meanings is obviously in line with current mainstream practice in WSD. The following comment particularly highlights the relevance and potential applicability of the model in our investigation of lexical sensitivity in WSD: "… It is possible that words rated low in context availability largely possess context-dependent knowledge which is relatively inaccessible when the words are presented in isolation. However, when such words are presented in supportive contexts, this context-dependent information becomes highly available for deriving meaning, eliminating potential differences in comprehension between abstract and concrete words." (Schwanenflugel, 1991: p.246)

Hence in the current study, we try to apply the context availability model in our investigation of the relationship between the effectiveness of various knowledge sources (in terms of the disambiguation performance) and the availability of characteristic linguistic context distinguishing one sense from the others for a particular target word. However, we will have to introduce a variation to the model. We have to distinguish between lexical and sense concreteness, the confusion of which is also a major inadequacy in psycholinguistic studies of the concreteness effect. On the one hand, the existence of polysemy means that a word can have multiple senses, but when psycholinguists attempt to norm the concreteness ratings from human subjects, there has been no control on how the subjects actually come up with a rating for the word as a whole. On the other hand, especially in view of the phenomena of sense extensions and metaphorical usages, polysemous words may consist of a mix of both concrete and abstract meanings, and it would make better sense to discuss the concreteness effect at the sense level instead of, or at least in addition to, the word level. This is particularly critical when word sense disambiguation is concerned.

We thus hypothesise that the differential effectiveness of individual knowledge sources is a result of the varied availability of characteristic linguistic context which serves to distinguish one sense from the others for a particular target word in the first place. This difference thus leads to different information susceptibility of individual target words, which is in turn reflected in the disambiguation performance, indirectly as the difficulty of WSD, giving rise to the long standing issue of lexical sensitivity.

3 The Current Study

We first set up a simple WSD experiment, running a supervised learning algorithm based on Support Vector Machines, with various knowledge sources (including topical contexts, local collocations, and local syntactic contexts) and their combinations on the noun samples in the SENSEVAL-3 English lexical sample task. The most frequent sense was used as the baseline. The disambiguation results were analysed and compared across individual target words. The algorithms implemented in the WEKA package (Hall et al., 2009), with all default settings, were used. For tokenisation and tagging of the data, the tokeniser and tagger available with the Lund University dependency parser (Johansson and Nugues, 2008) were used, although we did not use the parser specifically for this study.

3.1 Dataset

The data available for target nouns tested in the SENSEVAL-3 English lexical sample task were used. According to Mihalcea et al. (2004), the examples were extracted from the British National Corpus and the sense annotation was done using the Open Mind Word Expert system (Chklovski and Mihalcea 2002), and the sense inventory used for the nouns was WordNet 1.7.1 (Miller, 1995). Table 1 shows the target nouns with the number of senses and the distribution of concrete and abstract senses, as well as the number of training and testing instances for each noun. There are 20 items, with 3 to 9 senses, averaging at 5.35 senses.[1] The number of training examples for each sense varies considerably. The concrete/abstract classification of the senses was based on the lexicographer files in WordNet. Senses are organised under 45 lexicographer files based on syntactic category and logical groupings, and 26 of them are relevant to noun senses. We considered 7 of them concrete classes and the remaining 19 abstract classes. The concrete classes thus include *animal*, *artifact*, *body*, *food*, *object*, *person*, and *plant*. The abstract classes are *act*, *attribute*, *cognition*, *communication*, *event*, *feeling*, *group*, *location*, *motive*, *phenomenon*, *possession*, *process*, *quantity*, *relation*, *shape*, *state*, *substance*, *time*, and *Tops* (the unique beginner for nouns).

[1] These only cover the senses with training examples, not all senses listed in the sense inventory, hence the slight difference from the figures stated in Mihalcea et al. (2004).

3.2 Knowledge Sources

In this study, we focus on three types of disambiguating information: topical contexts, local collocations, and shallow syntactic information. They are realised in the form of bag of words, single words and word combinations in surrounding context, and the POS n-grams of neighbouring words, respectively, as binary features for the learning algorithm.

Topical Contexts (TC)

Topical contexts capture the broad conceptually related words, which are expected to reflect the topic or domain in which a sense often occurs. For this study we collected from the training examples all the noun and verb lemmas within a window of ±50 words from the target as features. Then in each testing instance, if any of those lemmas are found in a window of ±50 words from the target, the corresponding feature will have value 1, otherwise 0.

Local Collocations (LC)

The collocation patterns were approximated by the lemma unigrams, bigrams and trigrams in the local context of the target word, within a window of ±3 words. From the training instances, unigrams w_{-3}, w_{-2}, w_{-1}, w_1, w_2, and w_3, bigrams $w_{-3}w_{-2}$, $w_{-2}w_{-1}$, w_1w_2, and w_2w_3, and trigrams $w_{-3}w_{-2}w_{-1}$, $w_{-1}w_0w_1$, and $w_1w_2w_3$, were extracted as features. The word form of the target word was also included.

Shallow Syntactic Information (SS)

For this knowledge source, we collected features from the POS n-grams of the neighbouring words and the target word itself in the training instances, namely p_{-3}, p_{-2}, p_{-1}, p_0, p_1, p_2, p_3, $p_{-3}p_{-2}$, $p_{-2}p_{-1}$, p_1p_2, p_2p_3, $p_{-3}p_{-2}p_{-1}$, $p_{-1}p_0p_1$, and $p_1p_2p_3$.

3.3 Procedures

WSD results were first obtained with individual classifiers using various combinations of the knowledge sources. The results were then subject to comparison and error analysis, with respect to the intra-POS variation for the effectiveness of different knowledge sources.

3.4 Results and Analysis

As seen from Table 1, the target words have considerably different number of training and testing instances. Moreover, most of them are abstract. Of the 20 items, 9 only have abstract senses, and the rest have a mix of concrete and abstract senses. None is entirely concrete. Among the 107 senses for all words, only 24 are concrete senses. So the data is in some way biased in their concreteness. Although running WSD experiments on SENSEVAL data allows better comparison with previous studies, ideally there should be better control over the concreteness distribution especially for the purpose of this investigation. For this study, we will just note this deficiency.

Target Word	Senses	Con	Abs	Train	Test
argument	5	0	5	221	111
arm	5	4	1	266	133
atmosphere	5	1	4	161	81
audience	4	0	4	200	100
bank	9	4	5	262	132
degree	7	0	7	256	128
difference	5	0	5	226	114
difficulty	4	0	4	46	23
disc	4	3	1	200	100
image	6	3	3	146	74
interest	7	0	7	185	93
judgment	7	0	7	62	32
organization	4	0	4	112	56
paper	7	1	6	232	117
party	5	1	4	230	116
performance	5	0	5	172	87
plan	3	1	2	166	84
shelter	4	2	2	196	98
sort	4	1	3	190	96
source	7	3	4	64	32

Table 1: Sense distribution and data size

Table 2 shows the results from the various classifiers with different knowledge sources (TC for Topical Contexts, LC for Local Collocations, SS for Shallow Syntactic Information, ALL for the combination of the above, and Base is the baseline from the most frequent sense). The figures refer to precision, which is the same as recall in this case since coverage is 100% for all target words.

Most results in Table 2 are above the baseline. However, contrary to what most previous studies might have observed, especially if we look at individual target words, combining all knowledge sources does not necessarily give the best result. Hence the overall scores may sometimes be misleading as to the effectiveness of various knowledge sources to individual target words. It can be seen that the accuracy varies across different target words. For instance, using all knowledge sources, the result ranges from 0.391 for "difficulty" to 0.881 for "plan". The number of training instances available may make a difference, but for contrasting cases like "performance" and "plan" in this study, something else must be responsible for the different levels of difficulty as is apparent in the disambiguation results.

Target Word	TC	LC	SS	ALL	Base
argument	0.486	0.532	0.486	0.505	0.514
arm	0.850	0.872	0.857	0.865	0.820
atmosphere	0.716	0.667	0.580	0.679	0.667
audience	0.750	0.820	0.710	0.800	0.670
bank	0.841	0.765	0.614	0.818	0.674
degree	0.734	0.797	0.648	0.773	0.609
difference	0.474	0.518	0.447	0.623	0.404
difficulty	0.348	0.478	0.261	0.391	0.174
disc	0.780	0.480	0.420	0.710	0.380
image	0.595	0.608	0.419	0.649	0.365
interest	0.570	0.667	0.656	0.731	0.419
judgment	0.563	0.344	0.313	0.531	0.281
organization	0.768	0.768	0.643	0.768	0.732
paper	0.504	0.513	0.462	0.632	0.256
party	0.759	0.664	0.552	0.741	0.621
performance	0.506	0.322	0.322	0.425	0.264
plan	0.845	0.833	0.774	0.881	0.821
shelter	0.551	0.653	0.582	0.643	0.449
sort	0.646	0.719	0.688	0.698	0.656
source	0.688	0.563	0.406	0.625	0.656
Overall	0.666	0.652	0.572	0.698	0.542

Table 2: Disambiguation results

Regarding the concreteness effect, Table 3 shows the overall results with all knowledge sources and the baselines with respect to the concreteness of the senses for the words. Although the SENSEVAL-3 data contain more words with only abstract senses, the results apparently suggest that words with only abstract senses are more difficult to disambiguate than those with a mix of concrete and abstract senses, as is evident from the lower scores for the former in general.

Considering the effectiveness of various knowledge sources on individual target words, words with entirely abstract senses are apparently more susceptible to local features in addition to topical features. For instance, LC alone already gives better results than TC for 5 of the 9 target nouns with only abstract senses, compared to 6 of 11 words with a mix of abstract and concrete senses showing the same trend, not to mention that many of the nouns in the latter group actually consist of more abstract senses than concrete senses. In addition, with the addition of local features, only 2 out of 9 nouns with only abstract senses suffered a drop in the final score, compared to 5 out of 11 nouns with a mix of concrete and abstract senses were adversely affected. Topical contexts have usually been found to work well for nouns, but obviously their advantage is not as apparent in this study in the presence of predominantly abstract senses for the target nouns.

Concreteness	Baseline	SVM (All)
Only abstract senses	0.489	0.645
Both abstract and concrete	0.579	0.734
Overall	0.542	0.698

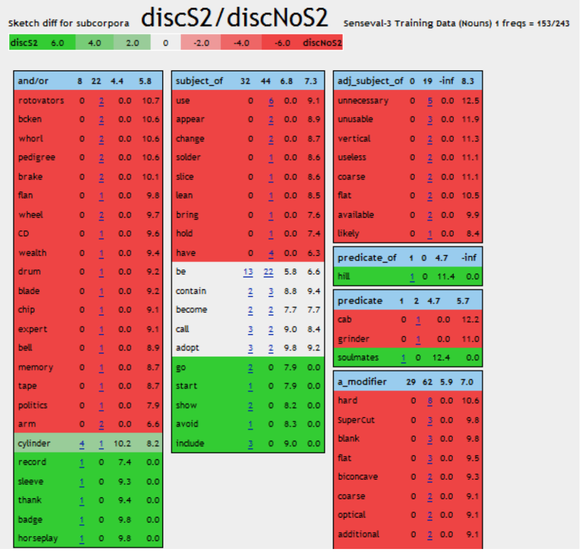
Table 3: WSD results w.r.t. concreteness

As mentioned, we will attempt to explain for the disambiguation results on concrete and abstract senses from the perspective of context availability. To this end, we consider the sense distinctions from the lexicographers' perspective.

Lexicographers distinguish senses by many criteria, most notably including: syntactic patterns, collocation patterns, colligation patterns, and domain. If one considers senses the artifacts from lexicography (e.g. Kilgarriff, 2006), it makes sense to think about WSD from lexicographers' point of view, because whether they rely on sufficient characteristic contextual difference to distinguish the senses to start with will directly affect the difficulty of subsequent disambiguation and the usefulness of various knowledge sources for this purpose. Hence we try to assess context availability with the Sketch Engine, an important tool for computational lexicography.

The Sketch Engine is a corpus query system widely used in modern computational corpus-based lexicography. It takes as input a corpus of any language and a corresponding set of grammar patterns, and generates word sketches for the words of that language; whereas word sketches are one-page automatic, corpus-based summaries of a word's grammatical and collocational behaviour (Kilgarriff et al., 2004). Sketch difference is also one of the many functions available in the Sketch Engine. It provides useful summaries in how pairs of near-synonyms differ, allowing users to compare and contrast the grammatical and collocational patterns of two words with apparently similar meanings.

We take advantage of the sketch difference function for comparing and contrasting individual senses of a word, to identify important grammatical and collocational patterns within specific grammatical relations critical for their distinction. To do this, we created sense sub-corpora for the Sketch Engine. All examples were extracted from the training data and stored in different files according to individual senses. A corpus was created in Sketch Engine, treating each set of examples as a sub-corpus, and all other senses of the same word as another sub-corpus, to facilitate subsequent comparison of prominent contexts among senses. For each target noun, we obtained the sketch difference for each of its senses with the rest of its senses, and analysed for common patterns and unique patterns with respect to sense concreteness and difficulty of WSD. For the word sketch patterns, we used the default English Penn Treebank sketch grammar available from the Sketch Engine. Typical grammatical relations specified in the word sketch patterns relevant to nouns include object_of (indicating the verbs which usually take the noun as object), a_modifier (indicating the adjectival pre-modifier for the noun), pp_%s (indicating common prepositional phrases following the noun), etc.



Sketch diff for subcorpora discS2/discNoS2 Senseval-3 Training Data (Nouns) 1 freqs = 153/243

discS2 6.0	4.0	2.0	0	-2.0	-4.0	-6.0 discNoS2

and/or	8	22	4.4	5.8
rotovators	0	2	0.0	10.7
bcken	0	2	0.0	10.6
whorl	0	2	0.0	10.6
pedigree	0	2	0.0	10.6
brake	0	2	0.0	10.1
flan	0	1	0.0	9.8
wheel	0	2	0.0	9.7
CD	0	1	0.0	9.6
wealth	0	1	0.0	9.4
drum	0	1	0.0	9.2
blade	0	1	0.0	9.2
chip	0	1	0.0	9.1
expert	0	1	0.0	9.1
bell	0	1	0.0	8.9
memory	0	1	0.0	8.7
tape	0	1	0.0	8.7
politics	0	1	0.0	7.9
arm	0	2	0.0	6.6
cylinder	4	1	10.2	8.2
record	1	0	7.4	0.0
sleeve	1	0	9.3	0.0
thank	1	0	9.4	0.0
badge	1	0	9.8	0.0
horseplay	1	0	9.8	0.0

subject_of	32	44	6.8	7.3
use	0	6	0.0	9.1
appear	0	2	0.0	8.9
change	0	2	0.0	8.7
solder	0	1	0.0	8.6
slice	0	1	0.0	8.6
lean	0	1	0.0	8.5
bring	0	1	0.0	7.6
hold	0	1	0.0	7.4
have	0	4	0.0	6.3
be	13	22	5.8	6.6
contain	2	3	8.8	9.4
become	2	2	7.7	7.7
call	3	2	9.0	8.4
adopt	3	2	9.8	9.2
go	2	0	7.9	0.0
start	1	0	7.9	0.0
show	2	0	8.2	0.0
avoid	1	0	8.3	0.0
include	3	0	9.0	0.0

adj_subject_of	0	19	-inf	8.3
unnecessary	0	5	0.0	12.5
unusable	0	3	0.0	11.9
vertical	0	2	0.0	11.3
useless	0	2	0.0	11.1
coarse	0	2	0.0	11.1
flat	0	2	0.0	10.5
available	0	2	0.0	9.9
likely	0	1	0.0	8.4

predicate_of	1	0	4.7	-inf
hill	1	0	11.4	0.0

predicate	1	2	4.7	5.7
cab	0	1	0.0	12.2
grinder	0	1	0.0	11.0
soulmates	1	0	12.4	0.0

a_modifier	29	62	5.9	7.0
hard	0	8	0.0	10.6
SuperCut	0	3	0.0	9.8
blank	0	3	0.0	9.8
flat	0	3	0.0	9.5
biconcave	0	2	0.0	9.3
coarse	0	2	0.0	9.1
optical	0	2	0.0	9.1
additional	0	2	0.0	9.1

Figure 1: Example of Sketch Difference

Figure 1 shows an example of the sketch differences between the second sense of the target noun “disc” (phonograph record) and its other senses (circular plate / magnetic disk / saucer) displayed by the Sketch Engine.

Let us illustrate our analysis with two examples. For instance, all senses for “degree” are abstract, as shown below. Table 4 shows a partial confusion matrix when TC and LC are used respectively.

1: [Attribute] {degree, grade, level} – a position on a scale of intensity or amount or quality

2: [Attribute] {degree} – the seriousness of something

3: [Cognition] {degree} – the highest power of a term or variable

4: [Communication] {academic degree, degree} – an award conferred by a college or university signifying that the recipient has satisfactorily completed a course of study

5: [Quantity] {degree, arcdegree} – a measure for arcs and angles

6: [Quantity] {degree} – a unit of temperature on a specified scale

7: [State] {degree, level, stage, point} – a specific identifiable position in a continuum or series or especially in a process

Expected \ Predicted	1	4	7
1	TC: 76 LC: 74	TC: 2 LC: 4	--
4	TC: 13 LC: 2	TC: 16 LC: 27	--
7	TC: 10 LC: 11	TC: 1 LC: 0	--

Table 4: Partial confusion matrix for “degree”

For the “degree” example, only Sense 1, 4 and 7 could be considered to have a reasonable number of training examples. Looking at the performance with TC and LC respectively, obviously Sense 7 is the most difficult because neither knowledge source was able to get any of the Sense 7 test instances correct. The confusion between Sense 1 and Sense 4 is obvious, and it is apparent that the use of local collocations is very effective to tell apart Sense 4 from Sense 1. The sketch differences show that Sense 1 has a lot of common patterns with non-Sense 1 data. However, it is the most frequent sense and might therefore have an advantage. On the other hand, Sense 4 has few common patterns with other senses but has considerable distinct patterns of its own with regard to local collocation and syntactic relations. Sense 7, however, shares many common patterns with other senses, but only has a few distinct yet not so characteristic patterns. This probably explains the benefits of adding local features for reducing the errors for Sense 4, as well as its lack of effect on disambiguating for Sense 7.

Turning to an example of mixed-sense target word, local features are destructive for "disc". The senses for the word are listed below. Sense 1 to Sense 3 are concrete, and Sense 4 is abstract. Table 5 shows the confusion matrix when TC and LC are used respectively.

1: [Artifact] {disk, disc} – a thin flat circular plate

2: [Artifact] {phonograph record, phonograph recording, record, disk, disc, platter} – sound recording consisting of a disc with continuous grooves; formerly used to reproduce music by rotating while a phonograph needle tracked in the grooves

3: [Artifact] {magnetic disk, magnetic disc, disk, disc} – (computer science) a memory device consisting of a flat disk covered with a magnetic coating on which information is stored

4: [Shape] {disk, disc, saucer} – something with a round shape like a flat circular plate

For the "disc" example, the impact of availability of training instances can be considered insignificant, as all four senses have over 30 instances. From Table 5, obviously TC is a much more effective knowledge source, at least for distinguishing among Senses 1 to 3. The sketch differences show that Sense 1 shares relatively many common patterns with non-Sense 1 data, and so does Sense 4. Sense 2 and Sense 3, on the other hand, share fewer common patterns with others. This possibly predicts the confusability between Sense 1 and Sense 4. Moreover, the unique patterns for individual senses are still restricted to the collocation patterns within particular grammatical relations, instead of any sense enjoying a unique syntactic pattern not found in others. This could explain why features based on words and lemmas are more effective for disambiguating this word, while the addition of local syntactic information does not help at all.

Expected \ Predicted	1	2	3	4
1	TC: 24 LC: 11	TC: 1 LC: 13	TC: 2 LC: 3	--
2	TC: 1 LC: 3	TC: 37 LC: 32	TC: 0 LC: 2	TC: 0 LC: 1
3	TC: 9 LC: 7	TC: 0 LC: 12	TC: 15 LC: 5	--
4	TC: 6 LC: 6	TC: 3 LC: 4	TC: 0 LC: 1	TC: 2 LC: 0

Table 5: Confusion matrix for "disc"

3.5 Implications on Lexical Sensitivity

From the above analysis, we have observed the following: First, nouns with only abstract senses are relatively more difficult to disambiguate than those with a mix of abstract and concrete senses, as seen from the overall scores for the two kinds of words. Second, the addition of local collocation and syntactic information to topical contexts often improves the overall score, but the actual effect varies across individual target words. Some benefit more from the combined features while others may suffer a drop in the final scores. Third, local collocation and syntactic features seem to play a more significant role on the disambiguation of abstract senses than concrete senses.

Past studies have observed that in general adding topical or bag-of-word features is more beneficial for nouns whereas adding local and collocational features works better for verbs and adjectives, but as we have observed in this study, such advantages do not necessarily apply to all words (and their senses) in the whole syntactic category. This means that POS alone may not be adequate to account for the lexical sensitivity of WSD, especially in view of the intra-POS variation with respect to individual knowledge sources. The common property shared by instances which can be effectively disambiguated by a certain kind of knowledge source or contextual feature is, simply speaking, context availability and the linguistic properties used by lexicographers for their distinction in the sense inventory in the first place.

The POS effect observed in previous studies could thus be understood this way. There are typical syntactic contexts in which words of different POS are bound to occur. For instance, nouns are often used in the subject and object positions and thus whether we find a verb before or after the target noun or whether its previous word is a determiner may not be a very good contextual feature in general because the various senses of a given noun may all occur in such similar contexts. On the contrary, if one sense of the noun tends to appear in very specific constructions, such as in very unique prepositional phrases, then in such cases one can expect local collocations and n-gram combinations to be relatively useful for distinguishing this sense from the others. An illustrative example is the target word "audience", as one of its senses is based on the specific usage of “the rights of audience”, which accounts for the particular effectiveness of LC and SS. Thus one problem with previous findings on the relation between knowledge sources and POS is that it may be too crude to look at lexical sensitivity in terms of POS alone and from the overall disambiguation scores, as the precise effect on individual words could vary considerably. For example, for the intra-POS variations among nouns, in this study we have observed the concreteness effect. The analysis suggested that concrete senses tend to rely more on topical information or they are more often used in distinctively different domains, while abstract senses are more likely to be characterised by their special local contexts such as the occurrence in particular PP or followed by particular PP, in addition to the topic or domain in which they are often used. The impact of sense concreteness, after all, is coupled with the actual context availability of individual senses, which affects the ease of disambiguation and the effectiveness of various knowledge sources. The model will thus predict that while sense dispersion or granularity will affect the difficulty of disambiguation, but if sufficient characteristic contexts can be associated with the senses and such contexts exist in the data, even closely related senses (such as an originally concrete sense and its abstract and metaphorical extension) could still be effectively disambiguated with the relevant knowledge sources.

4 Conclusion and Future Directions

While many previous studies have demonstrated the benefits or disadvantages of using certain knowledge sources for words of particular POS, in the current study we further address the intra-POS variations and discuss lexical sensitivity with respect to sense concreteness. As the context availability model in psycholinguistics predicts, although concrete words are more easily understood than abstract words, the concreteness effect will disappear if the stimuli were controlled for the ease to come up with an associative context.

Our analysis of WSD results on the noun samples in the SENSEVAL-3 English lexical sample task has allowed us to observe that words with only abstract senses tend to have lower disambiguation scores and are thus more difficult than those with a mix of abstract and concrete senses. Moreover, the benefit of adding local contextual information to topical contexts in disambiguation varies across target words, and it depends on the context availability of individual senses and the basis by which lexicographers distinguish and characterise them in the first place. These observations shed further light on the lexical sensitivity issue. In addition to factors like POS, sense granularity, number of senses, availability of training samples, etc., there is something about the intrinsic nature of individual words, such as concreteness, which may affect their susceptibility to different knowledge sources in disambiguation. It is therefore more appropriate to consider the lexical sensitivity in WSD in terms of information susceptibility, which depends on how the senses of the words were distinguished in the first place and whether their typical contexts are characteristic enough and available in most instances, resulting in the differential effectiveness of individual knowledge sources on different target words. To this end, WSD might be treated as the reverse engineering of lexicography, especially if one accepts that senses are the artifacts from lexicography. In this way, the selection of features and their combinations and weighting with specific learning algorithms could be made genuinely sensitive to individual lexical items.

For future work, we plan to deepen our investigation, making more systematic use of tools like the Sketch Engine to quantify context availability and to predict the usefulness of individual knowledge sources for WSD; and extend our testing and analysis to verbs and adjectives, to give a fuller picture of lexical sensitivity across different parts-of-speech.

Acknowledgments

The work described in this paper was supported by grants from the Department of Chinese, Translation and Linguistics of the City University of Hong Kong.

References

Agirre, E., & Stevenson, M. 2006. Knowledge sources for WSD. In E. Agirre, & P. Edmonds (Eds.), *Word Sense Disambiguation: Algorithms and Applications*. Dordrecht, The Netherlands: Springer.

Bleasdale, F.A. 1987. Concreteness dependent associative priming: Separate lexical organization for concrete and abstract words. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13,* 582-594.

Chklovski, T., & Mihalcea, R. 2002. Building a sense tagged corpus with Open Mind Word Expert. In *Proceedings of the Workshop on Word Sense Disambiguation: Recent Successes and Future Directions,* Philadelphia. pp.116-123.

Chugur, I., Gonzalo, J., & Verdejo, F. 2002. Polysemy and Sense Proximity in the Senseval-2 Test Suite. In *Proceedings of the Workshop on Word Sense Disambiguation: Recent Successes and Future Directions,* Philadelphia. pp.32-39.

Edmonds, P., & Cotton, S. 2001. SENSEVAL-2: Overview. In *Proceedings of the Second International Workshop on Evaluating Word Sense Disambiguation Systems (SENSEVAL-2),* Toulouse, France. pp.1-6.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H. 2009. The WEKA Data Mining Software: An Update. *SIGKDD Explorations, Volume 11, Issue 1.*

Johansson, R., & Nugues, P. 2008. Dependency-based syntactic-semantic analysis with PropBank and NomBank. In *Proceedings of the 12th Conference on Computational Natural Language Learning (CoNLL 2008),* Manchester. pp.183-187.

Jorgensen, J. 1990. The psychological reality of word senses. *Journal of Psycholinguistic Research, 19,* 167-190.

Kilgarriff, A. 2006. Word Senses. In E. Agirre, & P. Edmonds (Eds.), *Word Sense Disambiguation: Algorithms and Applications*. Dordrecht, The Netherlands: Springer.

Kilgarriff, A., & Rosenzweig, J. 1999. English SENSEVAL: Reports and results. In *Proceedings of the 5th Natural Language Processing Pacific Rim Symposium (NLPRS '99),* Beijing, China.

Kilgarriff, A., Rychly, P., Smrz, P., and Tugwell, D. 2004. The Sketch Engine. In *Proceedings of EURALEX 2004.*

Krahmer, E. 2010. What Computational Linguists Can Learn from Psychologists (and Vice Versa). *Computational Linguistics, 36(2),* 285-294.

Kroll, J.F., & Merves, J.S. 1986. Lexical access for concrete and abstract words. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 12,* 92-107.

Kwong, O.Y. 2008. A preliminary study on the impact of lexical concreteness on word sense disambiguation. In *Proceedings of the 22nd Pacific Asia Conference on Language, Information and Computation (PACLIC 22),* Cebu, Philippines. pp.235-244.

Kwong, O.Y. 2012. *New Perspectives on Computational and Cognitive Strategies for Word Sense Disambiguation*. Springer Briefs in Speech Technology. Springer.

Màrquez, L., Escudero, G., Martínez, D., & Rigau, G. 2006. Supervised corpus-based methods for WSD. In E. Agirre, & P. Edmonds (Eds.), *Word Sense Disambiguation: Algorithms and Applications*. Dordrecht, The Netherlands: Springer.

Mihalcea, R. 2002. Word sense disambiguation with pattern learning and automatic feature selection. *Natural Language Engineering, 8(4),* 343-358.

Mihalcea, R., Chklovski, T., & Kilgarriff, A. 2004. The SENSEVAL-3 English Lexical Sample Task. In *Proceedings of the 3rd International Workshop on the Evaluation of Systems for the Semantic Analysis of Text (SENSEVAL-3),* Barcelona, Spain. pp.25-28.

Miller, G. 1995. WordNet: A lexical database. *Communication of the ACM, 38(11)*, 39-41.

Paivio, A., Yuille, J.C., & Madigan, S.A. 1968. Concreteness, imagery, and meaningfulness values for 925 nouns. *Journal of Experiment Psychology, Monograph Supplement, 76(1, Pt.2)*, 1-25.

Pedersen, T. 2002. Assessing system agreement and instance difficulty in the lexical sample tasks of SENSEVAL-2. In *Proceedings of the Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*, Philadelphia, PA, USA. pp.40-46.

Resnik, P., & Yarowsky, D. 1999. Distinguishing systems and distinguishing senses: new evaluation methods for word sense disambiguation. *Natural Language Engineering, 5(2)*, 113-133.

Schwanenflugel, P.J. 1991. Why are abstract concepts hard to understand? In P.J. Schwanenflugel (Ed.), *The Psychology of Word Meanings*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Schwanenflugel, P.J., Harnishfeger, K.K., & Stowe, R.W. 1988. Context availability and lexical decisions for abstract and concrete words. *Journal of Memory and Language, 27*, 499-520.

Taft, M. 1991. *Reading and the Mental Lexicon*. Hove, East Sussex: Lawrence Erlbaum Associates.

Yarowsky, D., & Florian, R. 2002. Evaluating sense disambiguation across diverse parameter spaces. *Natural Language Engineering, 8(4)*, 293-310.

Yuret, D., & Yatbaz, M.A., 2010. The noisy channel model for unsupervised word sense disambiguation. *Computational Linguistics, 36(1)*, 111-127.