

MSDA: Wordsense Discrimination Using Context Vectors and Attributes

Abdulrahman Almuhareb¹ and Massimo Poesio^{1,2}

Abstract. We present MSDA (Major Senses Discovery Algorithm) – a development over the context vector approach to (noun) sense discrimination [20, 24] that uses attributes and values instead of word features to cluster contexts, and does not require for the number of senses to be fixed beforehand. The algorithm achieves a precision of 89% on a dataset including both ambiguous and non-ambiguous nouns, twice that of previous algorithms.

1 INTRODUCTION

Arguably, the most difficult part of the task of acquiring lexical and ontological knowledge from text [2, 6-8, 13, 16, 18, 23, 24] is the identification of a word’s senses—i.e. trying to discover that a word such as *palm* can be used to express both the concept “the inner surface of the hand” (WordNet 2 sense 1), henceforth **palm1**, and the concept “any plant of the family *Palmae*” (WordNet 2 sense 3), henceforth **palm3** [5, 16, 20, 22, 24]. Not the least among the difficulties is the fact that whereas **major sense** distinctions like the one just mentioned can generally be made reliably, other distinctions—e.g. that between **palm1** above and the derived sense of “a linear unit based on the length or width of a human hand” (WordNet 2 sense 2) are more difficult [12].

Our own work on the acquisition of lexical knowledge has been motivated by the consideration that knowing about concept **attributes**³—e.g. knowing that **ships** have **captains**—is an essential aspect of the rep-

resentation of concepts in AI [4, 26], Linguistics [21] and Psychology [15]. We developed methods for extracting from the Web candidate concept attributes, and semantic classifier to select those that fit into a linguistically and philosophically motivated classification of attributes, finding that this leads to improvements at concept clustering.

The motivation for the work discussed in the present paper is the intuition that such information about attributes ought to help with sense discrimination as well, in that different senses of a word ought to be associated with different attributes. Thus, for example, instances of **palm1** are more likely to be connected with a **wrist** than instances of **palm3**; conversely, the latter will be more likely to have **coconuts**. (Of course, there will also be attributes in common. E.g. instances of both **palm1** and **palm3**, being physical objects, will have a length—although we conjectured that these would not be the most distinctive attributes.) In addition, we explicitly focus on the task of identifying major senses instead of all sense distinctions as identified, say, in WordNet.

The structure of the paper is as follows. After a brief overview of relevant literature on sense discrimination and on using attributes for concept clustering, we introduce our new algorithm for sense discrimination, MSDA, and present the experiment we used to evaluate it. Results and discussion follow.

2 BACKGROUND

2.1 Wordsense discrimination with context vectors

The old observation that context is required to determine a word’s meaning led, e.g. Miller and Charles [14] to suggest that similar meanings are often used in similar contexts. Schütze [24] implemented this intuition by extending his earlier vector space model [23] to identify senses of words by using two different types

¹ University of Essex, UK, email: aalmuh & poesio at essex.ac.uk.

² University of Trento, Italy.

³ The term **attribute** is used informally here to indicate the type of relational information about concepts that is expressed using so-called **roles** in Description Logics [4]—i.e. excluding **IS-A** style information (that cars are vehicles, for instance). It is meant to be a more restrictive term than the term **feature**, often used to indicate any property of concepts, particularly in Psychology. We are carrying out a systematic analysis of the sets of features used in work such as [25] (see Discussion).

of vectors: **first-order vectors** representing contexts in terms of neighboring words of the target word, as in his original model; and **second-order vectors**, used to represent a context as the centroid of the first-order vectors for the words occurring in that context—which amounts to using word vectors as features. The context vectors for a word are then clustered using agglomerative clustering in a predefined number of sense clusters. Schütze reported an accuracy of between 72% and 90%, but pointed out that his algorithm should be tested using more natural ambiguous words. Pedersen and Bruce [17] developed a second method for identifying senses, in which contexts were directly represented with first-order vectors, i.e. using words themselves as features. These first-order vectors were then clustered using agglomerative cluster and the EM algorithm.

Purandare and Pedersen [20] (henceforth: P&P) compared these two methods, also looking at different combinations of clustering and features, and concluded that first-order vectors and agglomerative clustering are the best choice when the sample data is large (e.g. about 4,000 contexts per word), while second-order vectors and the Repeated Bisections algorithm are the best when the sample data is small (e.g. around 50-200 contexts per word). P&P [20] also developed a (publicly available) software package called **SenseClusters** implementing a variety of context representation models include first-order and second-order vectors with and without dimensionality reductions. The software also provides several clustering algorithms including agglomerative, partitional, and Repeated Bisections [11] algorithms.

2.2 Mining concept attributes from the Web

As said earlier, most theories of concepts are based on the assumption that concepts are characterized in terms of semantic attributes or, more generally, features, but no previous work on the acquisition of lexical or ontological knowledge by concept clustering attempted to identify such semantic properties. In our own earlier work [2, 18] we developed methods for extracting candidate attributes from the Web using patterns as done, e.g. by Hearst [10] to identify **is-a** relations or Poesio et al. [19] to identify **part-of** relations. For example, Hearst [10] acquired information about hyponymy (**is-a** links) by searching for instances of patterns such as

NP { , NP } * or other NP

(as in, e.g. *bruises broken bones and other INJURIES*). In our own work, we used the pattern

"the * of the C [is|was]"

(suggested by e.g. Woods [26] as a test for ‘attribute-hood’) to search for candidate attributes of concept *C* in the Web (finding e.g. instances such as *the color of the car is*) and a separate pattern for identifying **values** of such attributes (such as *red*). We showed in [3] that using only such attributes and values results in better concept descriptions for the purposes of clustering than collecting all relations in which a concept occurs. We then trained a classifier using morphological information and information from the Web for filtering such candidate attributes by assigning them to one of five classes according to a scheme derived from the work of Pustejovsky [21] and Guarino [9]. We showed in [18] that a binary classifier filtering non-attributes this way (i.e. eliminating attributes not recognized as belonging to one of the five classes) leads to significant improvements in concept clustering.

The conjecture motivating the present work was that such methods for identifying attributes and values could work better as a feature selection mechanism than the methods used by Schütze or P&P for the purposes of identifying distinct context clusters, each representing a separate sense.

3 A NEW SENSE DISCRIMINATION ALGORITHM

Our new sense discrimination algorithm, MSDA (Major Senses Discovery Algorithm) improves over the algorithms implemented in SenseClusters in three respects: (i) it uses attributes and values as semantic features used to identify different senses; (ii) it does not require to predefine the number of senses; and (iii) it returns a flag specifying whether context information provides enough evidence to decide on a number of senses.

The pseudo-code for MSDA is shown in Figure 1. The algorithm begins by computing the set *C* of context vectors *c*. The algorithm is designed to work with large amounts of data such as those that can be extracted from the Web (we use up to 10,000 contexts per word, see below); hence, on the basis of the Purandare and Pedersen results, it uses first-order vectors, one for each occurrence of target noun *w*. After collecting the contexts, an iteration begins. Each loop consists of two phases. In **Phase 1**, the contexts in *C* are clustered into *n* clusters (at the beginning *n* = 2; it is incremented by 1 at each iteration). Against the recommendations of Purandare and Pedersen, we used Repeated Bisections for clustering, with a criterion function that tries to maximize the internal similarity of each cluster [11]. We did not use agglomerative clustering because we found that it produces clusters

with varying sizes: some clusters contain very large number of instances, while other ones a very small number, which is not adequate in our case because we are only looking for major senses, that should be frequent in the data.

Input: A set of contexts C ; the thresholds Φ_1 and Φ_2 ; and m , the number of features to be compared.

$n = 2$; $f \leftarrow \text{true}$;

repeat

Phase 1:
partition C in n clusters $k_1 \dots k_n$ by repeated bisections;

Phase 2:
foreach $i=1..n$ find the m most frequent attributes and values in cluster k_i ;
 $k_i, k_j \leftarrow$ the two clusters such that $\text{sim}(k_i, k_j)$ is max;
if $\text{sim}(k_i, k_j) \geq \Phi_2$ **then**
 $n \leftarrow n-1$; **halt**;
if $\Phi_1 < \text{sim}(k_i, k_j) < \Phi_2$ **then**
 $n \leftarrow n-1$; $f \leftarrow \text{false}$; **halt**;
 $n \leftarrow n+1$;

end repeat

Output: n , the number of the discovered senses; f , a flag indicating if MSDA is certain or not; and the top frequent attributes & values related to each sense.

Figure 1: Pseudo-code for MSDA.

In **Phase 2**, attributes and values are used to discriminate between clusters. They are extracted from the contexts in each cluster k_i using the methods developed in our previous work, and the top m most frequent attributes / values for each cluster are used to decide which clusters k_i and k_j are most similar (using the cosine similarity function on binary values). At this point, a decision is made whether to halt or to try to partition C into more clusters. The decision is made by comparing $\text{sim}(k_i, k_j)$ with two thresholds. If the similarity between the top m features of the two most similar clusters is \geq threshold Φ_2 , MSDA halts, and concludes that there are $(n - 1)$ major sense(s). If the similarity is less than or equals a second threshold Φ_1 ($\Phi_1 < \Phi_2$), n is incremented and the next iteration is attempted. If the similarity falls in between these two thresholds, the algorithm reports that it is uncertain about the results, and there are at least $(n - 1)$ major sense(s) for the target noun.

The intuition behind the decision procedure is as follows. The higher threshold Φ_2 is used to decide when two sets of attributes / values are similar enough that two senses should be considered the same. Suppose that noun w only has one sense. Then the two initial clusters should share many attributes / values, hence the similarity between them should be $\geq \Phi_2$. The lower threshold Φ_1 is used to decide when two clusters

share so few features that they belong to two different senses. Suppose noun w has two or more clearly distinct senses. Then the sets of most distinctive attributes should be very different, and the similarity should be $\leq \Phi_1$. The algorithm will try to find more distinctive senses (but it will return $n-1$ if it ends up splitting a cohesive cluster). In other words, MSDA continues partitioning the contexts until it finds two very similar clusters. Finally, if the similarity falls in between the two thresholds, the evidence is not strong enough to give a certain conclusion. The algorithm also halts reporting an uncertain conclusion if there are less than m distinct features for any of the clusters in the set.

MSDA compares valid attributes (e.g. qualities, activities, and parts) and valid values (e.g. *fast*, and *slow* for the quality *speed*) only; invalid attributes are filtered out using the 2-way attribute classifier proposed in [18], and invalid values are filtered out using methods described in [1]. Note also that to eliminate the effect of these semantic features on the clustering, these features are hidden in Phase 1, and only used in the comparison in Phase 2.

4 EVALUATION

We now discuss how we evaluated MSDA, and compared it with Purandare and Pedersen’s SC.

4.1 Test data

In order to facilitate the comparison with previous work, we tested MSDA using a combination of unambiguous, naturally ambiguous, and pseudo-ambiguous words as done by Schütze and others. Our natural words included two ambiguous nouns used by Schütze (**capital** and **interest**), the one used by Rapp [22] (**palm**), and three nouns that are unambiguous (according to WordNet) and frequent: *brigade*, *hydrogen*, and *ship*. These three unambiguous nouns were also used to create three artificially ambiguous words: *brigade_hydrogen_ship*, *hydrogen_ship*, and *brigade_hydrogen*. ‘Contexts’ for these words were created by replacing e.g. every occurrence of *hydrogen* and of *ship* with *hydrogen_ship*, as done by Schütze.

4.2 Data collection

We collected from the Web about 10,000 contexts of occurrence for each natural noun in the dataset. For the pseudo-ambiguous nouns, we included a balanced number of contexts for each of the natural nouns that make them: e.g. the contexts for *hydrogen_ship* contain 5,000 contexts related to *hydrogen* and 5,000 contexts related *ship*. The context vectors were built using

a window of size 50 around the target noun. In phase 2, attributes and values are extracted from each context using the patterns discussed in our previous work.

We experimented with different values for Φ_1 , Φ_2 , and m (the number of most frequent attributes / values for each cluster). The best results were obtained with $\Phi_1 = 0.15$, $\Phi_2 = 0.30$, and $m = 20$.

4.3 Comparisons

We also used the data collected as above to test SenseClusters on the same dataset using Purandare and Pedersen's [20] recommendations and choices for clustering a large data sample. Again, about 10,000 contexts were used for each noun. The context vectors were constructed using a scope of size 5 and a window of size 40 (i.e. 20 in either side of the target noun). As done by P&P, we fixed the number of clusters to 7, but discarded clusters with few instances, less than 2% of the total number of instances, the threshold used by P&P. (We also tried a higher threshold (20%) and obtained similar results.) First-order contexts were used, and clustered using the agglomerative algorithm. Less frequent neighboring words with overall frequency of less than 2 were ignored. Frequencies were weighted using the log-likelihood function using the threshold 3.841.

4.4 Results

The results using SenseClusters (SC) and MSDA are shown in Table 1. The second column lists the number of major senses found in the literature for the natural ambiguous nouns (and based on WordNet for the unambiguous nouns); the third column shows the number of senses found using SC; the fourth the number of senses found using MSDA.

Noun	Major Senses	SC	MSDA
brigade	1	1 ✓	1 ✓
hydrogen	1	3 ✗	1 ✓
ship	1	1 ✓	1 ✓
brigade_hydrogen	2	2 ✓	2 ✓
hydrogen_ship	2	3 ✗	2 ✓
brigade_hydrogen_ship	3	2 ✗	3 ✓
capital	2	1 ✗	2 ✓
interest	2	1 ✗	1 ✗
palm	2	2 ✓	2 ✓

Table 1: Results with SC and MSDA

As can be seen from the Table, MSDA works very well: it identifies the correct number of senses for 8 out of 9 nouns, 89%. By contrast, SenseClusters only

found the correct number of major senses for 4 out of 9 nouns, 44%.

4.5 Semantic features related to each sense

Table 2 shows the 4 most frequent semantic features (attributes and values) for each sense of the five ambiguous nouns found using MSDA. The features are ordered by frequency. Attribute features are preceded by '[a]', while value features are preceded by '[v]'. The first column in the table shows a manually assigned label for each sense.

Label	Top 4 Frequent Features
brigade_hydrogen	
brigade	[v]fire, [v]Texas, [v]1st, [a]commander
hydrogen	[v]liquid, [v]amide, [v]element, [v]fire
hydrogen_ship	
hydrogen	[v]liquid, [v]amide, [v]element, [v]molecular
ship	[a]captain, [a]name, [a]safety, [a]owner
brigade_hydrogen_ship	
brigade	[v]Texas, [v]1st, [v]3rd, [a]commander
hydrogen	[v]liquid, [v]amide, [v]element, [v]molecular
ship	[v]fire, [a]captain, [a]name, [a]side
capital	
assets	[v]fixed, [v]variable, [a]value, [v]circulating
district	[a]streets, [a]transfer, [a]removal, [a]heart
palm	
hand	[a]size, [v]right, [v]left, [a]base
tree	[v]date, [v]coconut, [v]oil, [v]sago
interest	
Common features: [a]extent, [a]holder, [a]transfer, [a]value, [v]controlling, [v]partnership, [v]security	

Table 2: Top 4 semantic features for each sense found by the MSDA algorithm.

Most of the attributes and values in the table are clearly related to the named sense. For example, the *assets* sense of *capital* includes the attribute *value* and the values *fixed*, *variable*, and *circulating*, which are clearly related to *assets*; while most of the attributes *streets*, *transfer*, *removal*, and *heart* are more related to the *district* sense of *capital* than to the *assets* sense.

The table also shows the data for the noun *interest* where MSDA failed to find the correct number of major senses: In the first iteration of Phase 2, MSDA found 7 common features (shown in the table) among the top 20 features of the two clusters of the *interest* contexts which gives a similarity score of 0.35 that is above the threshold Φ_2 .

5 DISCUSSION AND CONCLUSIONS

This work suggests using semantic features such as

attributes and values is not only a good way to build concept descriptions for the purposes of clustering, but also to discriminate between noun senses. This work however raises a number of questions that we hope to explore in further research, such as whether it is possible to come up with a clear enough repertoire of ‘major senses’ for some words that we will be able to assess whether in fact the algorithm is always successful at the task. (Unfortunately, the distinction between polysemy and homonymy is notoriously difficult to draw – see again Kilgarriff’s paper.)

ACKNOWLEDGMENTS

Abdulrahman Almuhareb is supported by King Abdulaziz City for Science and Technology (KACST), Riyadh, Saudi Arabia.

REFERENCES

- [1] Almuhareb, A. (2006) *Attributes In Lexical Acquisition*, PhD Thesis, Department of Computer Science, University of Essex.
- [2] Almuhareb, A. & Poesio, M. (2004) Attribute-based and value-based clustering: An evaluation. In *Proceedings of EMNLP*. Barcelona.
- [3] Almuhareb, A. & Poesio, M. (2005) Finding Attributes in the Web Using a Parser. In *Proceedings of Corpus Linguistics*. Birmingham.
- [4] Baader, F., Calvanese, D., McGuinness, D., Nardi, D. & Patel-Schneider, P. (Eds.) (2003) *The Description Logic Handbook: Theory, Implementation, and Applications*, Cambridge University Press.
- [5] Caraballo, S. A. (1999) Automatic construction of a hypernym-labeled noun hierarchy from text. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics*. College Park, MD, USA, 120-126.
- [6] Curran, J. R. & Moens, M. (2002) Improvements in Automatic Thesaurus Extraction. In *Proceedings of the Workshop on Unsupervised Lexical Acquisition*. Philadelphia, PA, USA, 59 - 67.
- [7] Dorow, B. & Widdows, D. (2003) Discovering Corpus-Specific Word Senses. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics*. Budapest, Hungary, 79-82.
- [8] Grefenstette, G. (1994) *Explorations in Automatic Thesaurus Discovery*, Boston, USA, Kluwer Academic Publishers.
- [9] Guarino, N. (1992) Concepts, attributes and arbitrary relations: some linguistic and ontological criteria for structuring knowledge bases. *Data and Knowledge Engineering*, 8, 249-261.
- [10] Hearst, M. A. (1998) Automated Discovery of Word-Net Relations. IN FELLBAUM, C. (Ed.) *WordNet: An Electronic Lexical Database*. MIT Press.
- [11] Karypis, G. (2003) *CLUTO: A clustering toolkit*. Technical Report #02-017 [online]. Department of Computer Science, University of Minnesota, Minneapolis, MN. Available from: <http://www-users.cs.umn.edu/~karypis/cluto/> [Accessed 27/09/2005].
- [12] Kilgarriff, A. (1997) I don't believe in word senses. *Computers and the Humanities*, 31(2), 91-113.
- [13] Lin, D. (1998) Automatic retrieval and clustering of similar words. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics*. Montreal, Canada, 768-774.
- [14] Miller, G. A. & Charles, W. G. (1991) Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1), 1-28.
- [15] Murphy, G. L. (2002) *The Big Book of Concepts*, The MIT Press.
- [16] Pantel, P. & Lin, D. (2002) Discovering word senses from text. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. Edmonton, Alberta, Canada, 613-619.
- [17] Pedersen, T. & Bruce, R. (1997) Distinguishing Word Senses in Untagged Text. In *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing*. Providence, RI, 197-207.
- [18] Poesio, M. & Almuhareb, A. (2005) Identifying Concept Attributes Using a Classifier. In *Proceedings of ACL Workshop on Deep Lexical Acquisition*. Ann Arbor, USA.
- [19] Poesio, M., Ishikawa, T., Walde, S. & Vieira, R. (2002) Acquiring lexical knowledge for anaphora resolution. In *Proceedings of LREC*. Las Palmas.
- [20] Purandare, A. & Pedersen, T. (2004) Word Sense Discrimination by Clustering Contexts in Vector and Similarity Spaces. In *Proceedings of the Conference on Computational Natural Language Learning (CoNLL)*. Boston, MA.
- [21] Pustejovsky, J. (1995) *The generative lexicon*, MIT Press.
- [22] Rapp, R. (2004) Utilizing the One-Sense-per-Discourse Constraint for Fully Unsupervised Word Sense Induction and Disambiguation. In *Proceedings of the fourth international conference on Language Resources and Evaluation (LREC 2004)*. Lisbon, Portugal.
- [23] Schütze, H. (1992) Dimensions of meaning. In *Proceedings of Supercomputing '92*. Minneapolis, 787-796.
- [24] Schütze, H. (1998) Automatic Word Sense Discrimination. *Computational Linguistics*, 24, 97-123.
- [25] Vinson, D. P., Vigliocco, G., Cappa, S. & Siri, S. (2003) The Breakdown of Semantic Knowledge: Insights from a Statistical Model of Meaning Representation. *Brain and Language*, 86(3), 347-365.
- [26] Woods, W. A. (1975) What's in a link: Foundations for semantic networks. IN BOBROW, D. G. & COLLINS, A. (Eds.) *Representation and understanding: studies in cognitive science*. New York, Academic Press.