EXPLOITING DISTRIBUTIONAL SIMILARITY FOR LEXICAL ACQUISITION

McCarthy Diana (diana@dianamccarthy.co.uk)

Lexical Computing Ltd., Brighton, East Sussex

Lexical acquisition has been dubbed the bottleneck of large scale robust natural language processing applications for at least two decades. There is now a substantial body of research dedicated to this important subfield of computational linguistics. Since the 1990s, researchers have turned to corpora for automatic lexical acquisition, rather than rely on extraction from existing online lexical resources. This allows for coverage of new domains, genres and languages without existing resources and where available resources do not provide sufficient coverage or require tailoring to the specific text type. A large body of lexical acquisition from corpora uses distributional similarity whereby the similarity between two words is calculated from the extent that the words have similar contexts of occurrence. Distributional similarity approaches are used for smoothing unseen events using data from seen events. They are also used as an approximation of semantic similarity since there is a strong tendency for words that exhibit similar distributional behaviour to share in their underlying semantics. This paper provides a summary of research that I, along with various collaborators, have conducted using distributional similarity to automatically acquire sense frequency information, selectional preferences and estimates of semantic non-compositionality of putative multiwords.

Key words: lexical acquisition, distributional similarity, NLP, semantic similarity

1. Introduction

Automatic lexical acquisition has received considerable interest for the past twenty years and more since without it computational linguistic systems simply will not scale and due to the emphasis on the lexicon as the appropriate repository for the majority of linguistic information (Gazdar, 1996). The focus quickly shifted from acquisition from electronic resources to acquisition from corpora since it was felt that this would avoid the errors and lack of coverage that beset man made resources. Extraction from corpora furthermore allows acquisition to languages and tailoring to domains which are not covered, or are poorly served by pre existing resources. Corpora also provide the much needed frequency information that is the backbone of computational linguistics systems, which since the 1990s are invariably statistical. That said, corpus approaches suffer from errors that arise in automatic processing of naturally occurring data and are dependent on sufficient language data of the appropriate type being available in electronic form. Neither approach provides a panacea and many solutions are found in hybrid approaches (Klavans and Resnik, 1996).

One major area of research in automatic acquisition from corpora has been the use of distributional similarity. In distributional similarity approaches, words are represented by the contexts that they occur in and the frequency of occurrence in these contexts. A vector capturing this information can be used directly for representation and a measure of distributional similarity is used to compare the representation of one word with that of another. Automatic distributional "thesauruses" can be produced from this data. In these thesauruses, a word entry is listed with other words that have the most distributional contexts in common with the target word. Distributional similarity can be applied to linguistic phrases beyond the lexical level (Mitchell and Lapata, 2008) however in this paper, we focus on the application to lexical acquisition.

Lexical acquisition encompasses a wide variety of different areas of linguistics: phonology, morphology, syntax and semantics. Certain aspects that relate to pragmatics are also beginning to receive attention, such as the widespread interest in sentiment. Topics that have been particularly prevalent have been the acquisition of word senses and information associated with specific senses such as collocations, subcategorisation (predicate argument structure), selectional restrictions or preferences (for parsing and semantic role labelling), and multiwords. In this paper I provide an overview of some of my research in lexical acquisition in the last decade focusing particularly on work exploiting distributional thesauruses. I will focus the paper on acquisition of word sense frequency information and non-compositionality detection of putative multiwords.

1.1. Distributional similarity

Distributional similarity is an approach which uses statistics concerning the contexts of occurrence of words and determines the similarity between two words given this information. A word is represented by a vector of values, usually frequency values, from a corpus and each dimension of the vector represents a particular context. The definition of context varies considerably. It can be a document, a specified grammatical relation or within a window of words around the target. Distributional similarity uses these vectors and calculates a similarity score designed to measure the similarity between the vectors. The distributional similarity score can be used for smoothing statistical models. In such an approach, seen information occurring with a word is used for a rarer or unseen word that is related to the more frequent word by distributional similarity. It can also used as an approximation of semantic similarity since there is a strong tendency for words that exhibit similar distributional behaviour to share in their underlying semantics. To this end the vectors have been used for semantic representation in vector space models (Schütze, 1998). The similarity score can also be used to produce a "distributional thesaurus" by ranking other words in terms of their similarity to the target word and the top K (where K is a threshold such as 10 or 50) words are provided in rank order as the nearest neighbours to the target word along with the distributional similarity score used to rank them.

There are many different distributional similarity measures (see Weeds (2003) for a survey). Though we have used various measures in our work (Weeds et al., 2004; McCarthy and Navigli, 2009), we have predominantly used the measure proposed by Lin (1998) and found it to perform well on our tasks. As a rule, we have used the grammatical relations output from RASP (Briscoe and Carroll, 2002) as our contexts, though we have also observed good results with proximity relations (McCarthy et. al., 2007) which bodes well for applying the methods to data without a suitable parser.

2. Word sense frequency acquisition

Words have different meanings and we expect our computational models to reflect this. Naturally, we therefore expect to represent different meanings in the lexicon somehow, and in doing so it is necessary to have an automatic method of associating the word forms in natural language data with the senses in the lexicon. Such automatic methods fall under the rubric of word sense disambiguation. Word sense frequency information is arguably the most important information for this enterprise.

Word sense disambiguation is performed using clues such as collocations and domain information which can be automatically acquired from training data where the target senses have been marked up by human annotators, or from existing resources, or automatically from corpora. The best performing word sense disambiguation methods however rely on a very simple heuristic to supplement information from the context. This is known as the first (or most frequent) sense heuristic. The first sense heuristic is particularly powerful (Navigli, 2009) and particularly so when the contextual evidence is weak and when the entropy is low, that is the sense frequency distribution for a given word is particularly skewed. Of course contextual evidence is required to disambiguate words effectively, nevertheless, in many typical texts there is a strong tendency for the same sense to occur throughout a discourse (Gale et al., 1992) . McCarthy et al. (2004) proposed a method to automatically determine the most likely sense given a particular corpus as training data and a predefined inventory of senses. Researchers had been using predominant sense information for many years but what was new in this work is that sense predominance could be estimated from corpus data that had not been annotated by hand. Manually tagging a corpus with word senses is a laborious and costly process (Ng, 1997). The use of an "unsupervised" system that did not require manually labelled training data meant that not only was the technique applicable to a language without a handtagged corpus (Iida et al., 2008), but also that the method can be applied to corpus data from a given domain which will give more appropriate sense frequency information compared to using a general purpose resource, at least for words that are salient to that domain (Koeling et al., 2005).

In this paper, we give a brief overview of the method and some of our main findings. For a full account of the method and results, please see McCarthy et al. (2007) and the various references in this paper.

2.1. Method

The approach first reported in McCarthy et al. (2004) works as follows. Given a listing of word senses from an inventory such as WordNet (1998), we calculate a ranking score over those senses. As an example, take the noun *tie*. In WordNet (version 3.0) there are in fact 9 senses, but if we use just the first three for this example we have

1. **necktie**, tie — (neckwear consisting of a long narrow piece of material worn (mostly by men) under a collar and tied in knot at the front; "he stood in front of the mirror tightening his necktie"; "he wore a vest and tie")
2. **affiliation**, association, tie, tie-up — (a social or business relationship; "a valuable financial affiliation"; "he was sorry he had to sever his ties with other members of the team"; "many close associations with England")
3. tie — (**equality of score** in a contest)

When we applied the Lin (1998) distributional similarity score to data from the British National Corpus (Leech, 1992) parsed with RASP, we observe the following top 10 neighbours with their corresponding distributional similarity scores used for ranking shown in parenthesis:

<u>BNC:</u>
links (0.165) shirt (0.162) scarf (0.152) jacket (0.142) bond (0.130) match (0.128) trousers (0.126) link (0.125) collar (0.125) dress (0.121)[1]

We can intuitively see that while the neighbours reflect different senses of *tie*, there are more that relate to the **necktie** sense. To calculate the ranking score for each sense, we take each distributional similarity score of each neighbour and allocate a proportion of it to each of the three senses. We do this such that the proportion is reflected in the semantic similarity between the sense and that neighbour. Neighbours are words and so may have multiple senses. To calculate the semantic similarity between a sense and a neighbour the algorithm picks whichever sense of the neighbour maximises the semantic similarity to the target word. The calculation for semantic similarity depends on what sense inventory we have. For our work with WordNet, we tried various measures from the WordNet Similarity Package (Patwardhan and Pedersen, 2003). The JCN (Jiang and Conrath, 1997) and Lesk (Lesk, 1986) proved to perform well. The JCN uses the hypernym structure of WordNet to estimate semantic similarity while Lesk uses the overlap of dictionary definitions. Lesk is therefore useful in many cases where a sense inventory is like a standard dictionary with definitions but without the semantic relationships encoded in WordNet. The method produces a score for each sense by summing the distributional similarity scores (0.165 0.162 etc.) each multiplied by a weight for that sense and that neighbour where the weight is the maximum semantic similarity (JCN for example) between that sense and any of the senses of that neighbour. Thus for a sense (s) of a word (w) the calculation is as follows:

[1] We use only the top 10 neighbours here for the sake of brevity.

$$ranking\ score(s \in senses(w)) = \sum_{i=1}^{k} distsim(w, n_i) \times maximum(jcn(ns \in senses(n_i)), s)$$

Where distsim represents the distributional similarity between w and the neighbour of w at rank i. jcn is the semantic similarity measure that weights the contribution from this neighbour according to its semantic similarity with s.

Thus, while there are neighbours obtained from the BNC related to different senses, the majority here are most strongly related (intuitively and by measuring with JCN) to the first **necktie** sense of *tie*.

Although we can get this information from sense tagged texts such as SemCor (Miller et al. 1993), the sense distributions will naturally vary in different domains. We can see this by looking at domain specific data we (Koeling et al., 2005) collected from the Reuters Corpus (Rose et al., 2002) in finance and sport. The top 10 neighbours of *tie* are:

<u>Finance:</u>
relation (0.329) links (0.247) relationship (0.232) cooperation (0.228) contact (0.142) partnership (0.141) trade (0.137) role (0.133) integration (0.133) finances (0.132)

<u>Sport:</u>
qualifier (0.191) match (0.174) clash (0.150) round (0.135) semifinal (0.132) series (0.129) fixture (0.125) matchup (0.120) encounter (0.120) win (0.116)

The majority of neighbours in Finance are most strongly associated with the **affiliation** sense of *tie*, whereas those from Sport are most strongly associated with the third **equality of score** sense.

2.2. Further work

In addition to the various studies with corpus data that has been classified for domain manually, we have also demonstrated that we can apply this method successfully where the corpus data needed for training our models has been marked up for domain automatically and also where the input data itself is likewise annotated automatically (Koeling et al., 2007) .

As well as adaptation to different domains, we (Iida et al., 2008) have also applied our method to another language, Japanese, and show that where a dictionary does not have the structure that WordNet does, then we can use the Lesk score. We also propose an adapted Lesk score which uses distributional similarity to refine the overlap measure between the definition of a sense to be ranked and the senses of the neighbours. Rather than summing the exact matches between any of the words occurring in the two definitions, we use the sum of the distributional similarity scores of the words in the paired definitions where words that are present in both get the maximum distributional similarity score of 1. This has the effect of coping with sparse data to give a more productive overlap method.

Another aspect of our more recent work is to use the sense ranking for word sense disambiguation, i. e. taking account the context rather than simply applying the top ranking sense irrespective of context. As well as automatically detecting the domain (Koeling et al., 2007; Koeling and McCarthy, 2007), we have used the ranking score to estimate the entropy of the sense distribution to better gauge when the predominant sense heuristic will be more powerful, because the distribution is skewed, or when the distribution is flatter and it is more important to look for contextual evidence (Jin et al., 2009). We have obtained modest improvements using the grammatical relation in the target sentence to help determine which neighbours, and therefore sense, is more relevant in the context (Koeling and McCarthy, 2008).We have also recently used the sense ranking information to help in initialising domain specific graphical methods for word sense disambiguation (Reddy et al., 2010). Accuracy improves by 11 percentage points when domain specific sense ranking information is used.

3. Non-compositionality detection of putative multiwords

A crucial aspect of lexical acquisition is to determine exactly which entries should be stored in the lexicon. Nevertheless, I and various collaborators have been developing acquisition methods that aim to detect cases of semantic non-compositionality because ultimately such techniques could be used to determine the boundaries of what entries go in the lexicon and what stays out.

Multiwords have received considerable attention in computational linguistics over the past decade and particularly since the seminal paper by Sag et al. (2002). There has been a series of ten workshops run at the main international computational linguistics conferences in the last decade focusing on various aspects of computational representation, handling and application of multiwords. One important and reoccurring issue is the difficulty of a precise definition to make a clear boundary between what is and what it not a multiword. Coverage of multiwords in man made lexicons varies considerably for this very reason and also because of their abundance and the fact that multiword neoglisms are coined all the time. There are many reasons why the boundaries vary, but for many purposes there is some level of idiosyncratic behaviour. This might be syntactic, for example *wine and dine,* or pragmatic, for example *good morning*[2], but in most cases we care about semantic non-compositionality which may give rise to other types of idiosyncratic behaviour. In addition to organisation of some of the multiword expression workshops and a journal special issue, my involvement in this area has been in automatic methods for detecting compositionality, or the lack of it, on the grounds that this will help determine the boundaries of what should be stored in the lexicon.

My research has focused on English and has emphasised the fact that compositionality is on a continuum. In McCarthy et al. (2003) we conducted experiments contrasting distributional similarity of the phrasals and the constituent verbs

[2] I am indebted to Timothy Baldwin for these examples.

to determine the extent that putative phrasal verbs (such as *blow up* and *eat up*) are compositional. In McCarthy et al. (2007) we conducted experiments on verb-object combinations, such as (*draw breath* and *light cigarette*). We again used distributional similarity, but this time rather than comparing the distributional profile of constituents to that of the whole phrase we used the nearest neighbours for modelling the selectional preference of the verb and then determine if the object was prototypical as an argument or not. If the object is not semantically related, using distributional similarity as a proxy for semantic similarity, to the typical objects seen with that verb then this is an indication of non-compositionality. We use preference strength directly to measure this. The next two subsections give a little more detail on these two works but we refer the interested reader to the papers cited for further details.

3.1. Detecting compositionality of phrasal verbs

For these experiments, we were interested in estimating the semantic compositionality of phrasal verbs which had been found by the RASP parser. We investigate various measures which compare the nearest neighbours of the verb constituent (e. g. *eat*) with the phrasal verb (e. g. *eat up*) or which scrutinize the list of nearest neighbours of the phrasal for occurrence of the constituents, or which combine both these approaches. More specifically the methods were:

- overlap: overlap of the top K neighbours of the phrasal and the constituent verb.[3]
- Sameparticle: the number of neighbours in the top 500 of the phrasal containing the same particle, for example *nibble up* has the same particle as *eat up*.
- Sampleparticle-simplex : as for Sameparticle but where we deduct the number containing the same particle which occurred in the simplex (constituent) verb's neighbours (the neighbours of *eat*).
- Simplexasneighbour: whether the simplex verb (*eat*) occurs in the top 50 nearest neighbours of the phrasal.
- Rankofsimplex: the rank of the simplex in the top 500 neighbours
- Scoreofsimple: the distributional similarity score of the simplex in the top 500 neighbours of the phrasal
- OverlapS: the overlap in the top K neighbours of the phrasal with those of the constituent verb's neighbours but where we remove all particles from the phrasal's neighbours (so for example, *nibble up* would become *nibble*).

We experimented with data from the BNC using Lin's measure of distributional similarity. We evaluated our methods by ranking a list of candidate phrasals according to these measures and correlating them using Spearman's rho with a gold-standard. We created the gold-standard by asking a set of three human annotators how compositional the candidate phrase was on a scale of 0–10 (idiomatic — fully compositional). Correlation was highest and highly significant for sameparticle, sameparticle-simplex and for the overlapS when using 30 or 50 neighbours. An interesting finding was that

[3] We experimented with different values of K, 30, 50, 100, 500.

statistics often used for multiword detection, such as Chi-squared, the log-likelihood ratio (Dunning, 1993) and pointwise mutual information (Church and Hanks, 1991) gave much lower, though significant, correlations. Phrasal frequency was not even significantly correlated.

3.2. Using Distributional Similarity for detecting compositionality of verb-object pairs

In this work, we used the dataset of verb-direct object pairs provided by Venkatapathy and Joshi (2005) which contained compositionality judgments on a scale following McCarthy et al (2003). This time, rather than use the distributional similarity neighbours for comparison of the constituents to the whole, we used them to build selectional preference models to estimate the preference strength of the verb for the given object. It is assumed that a weak preference for a particular direct object would indicate that the particular verb and object combination does not exhibit the normal semantic behaviour of the verb, and that this combination is non-compositional. For example, in the expressions *I'll eat my hat*, the direct object *hat* is not prototypical of the types of object we usually see with *eat* and our model should indicate this. We use a measure of selectional preference strength as an estimate of compositionality.

One issue for selectional preference acquisition is that it is acquired from automatically parsed data and multiwords are present in such data. Indeed selectional preference acquisition was one of our main motivations for detecting compositionality of multiwords in the first place (McCarthy et al. 2003). To avoid this problem we contrasted standard WordNet models (Li and Abe, 1998) which use direct object token instances from the training data to determine the classes, with type based models which use word types rather than tokens. We proposed both WordNet and distributional similarity type based models and contrasted these with the traditional token based models. Traditional token based models, such as (Resnik, 1993; Li and Abe, 1998) use direct object data for a given verb to populate the WordNet noun hierarchy with frequencies and obtain a probability distribution over WordNet classes. Our WordNet type based models use word types to determine the classes used for representation, rather than tokens, before then calculating the probability distribution using the tokens. Only if there are several types of a semantic class does the model include that class. For example, though *eat hat* might be reasonably frequent in a corpus, the type based models would not retain the probability under a **clothing** class simply because there are no other word types to support the use of that class in the model, whereas for *wear hat* that would not be the case due to the occurrence of words such as *coat*, *scarf* and *dress* which are semantically related. Furthermore, in these type based models we disambiguate an object that occurs at (directly or by virtue of hypernymy) several WordNet classes by assigning it to the class with the maximum number of types in the object data.

We contrasted the type based WordNet models with type based distributional models that use distributional similarity to group the objects in the training data into "classes". In these distributional similarity models the classes are a subset of the objects selected automatically so as to maximise[4] the inclusion of the object types from the training data for this verb in the top K neighbours. The training data for each verb was obtained from the direct objects detected for that verb from RASP parses of the BNC. The probability distribution associated with each class is then estimated using the frequency of the objects occurring as distributional neighbours (in the top K) to these words representing the classes. Where an object occurs as a neighbour of multiple words selected as classes, the class selected will be that which has the maximum number of object types as neighbours. A portion of the model acquired for the direct object slot of *park* is shown in table 1.

We compared these three types of model on the subset of the Venkatapathy and Joshi dataset that contained common nouns as objects (rather than adjectives, pronouns and complements). All models produced significant results. The type based WordNet models outperformed the token based models but were themselves outperformed by the models that used distributional similarity for the

Table 1. A portion of the distributional similarity selectional preference for the direct object of the verb 'park'

Class (probability)	Disambiguated objects (frequency)
van (0.86)	car (174) van (11) vehicle (8) . . .
mile (0.05)	street (5) distance (4) mile (1) . . .
yard (0.03)	corner (4) lane (3) door (1)

Classes without requiring a manually constructed resource like WordNet. This is an encouraging result as it means that the method can be applied to a language without such a resource.

The methods also outperformed the individual features that Venkatapathy had used on the same portion of the data. These features included vector space models (Baldwin et al., 2003), pointwise mutual information and an existing method for detecting compositionality using distributional similarity to find non productive combinations (Lin, 1999). The best results were obtained when using the distributional similarity selectional preferences combined with some of these other features. This demonstrates that while the selectional preferences are useful features for non-compositionality detection of verb-object multiwords, no one approach is a panacea.

3.3. Further work

We (Reddy et al., 2011) are currently engaged in further work to examine compositionality judgments of humans in more detail by considering not only judgments for the phrase as a whole, but also for the individual constituents. We are developing distributional similarity methods that likewise compare the distributional profile (the vector containing contexts of occurrence) of the constituent words with the vector for

[4] We use a greedy algorithm.

the whole phrase combined also with the distributional similarity between models of a composition of the constituent vectors and the vector for the whole phrase. The composition vector representations use both additional and multiplication composition functions over the constituent vectors (Mitchell and Lapata, 2008). Furthermore we refine the constituent vectors, inspired by Erk and Pado (2010) by considering only the contexts that are shared by both constituents but not including the contexts occurring with the candidate multiword.

4. Conclusions and future directions

In this paper, I have given a summary of research I have conducted, along with various collaborators, in exploiting distributional similarity for lexical acquisition. I have focused this paper on work on sense ranking and on compositionality detection for multiwords. The compositionality detection itself involved use of distributional similarity models for acquiring selectional preferences. There are others using distributional similarity for selectional preference acquisition (Erk, 2007) and we look forward to trying out these models for new purposes, such as automatic detection of diathesis alternations where previously we used token based WordNet models (McCarthy, 2000).

Another direction for research has been the representation of sense using distributional similarity. There are several strands of such research (Panel and Lin, 2002; Erk and McCarthy 2009). I am particularly interested in alternative ways of annotating and evaluating distributional models of semantics using paraphrases (McCarthy and Navigli, 2009; McCarthy et al., 2010), translations (Mihalcea et al. 2010) and usage similarity judgments (Erk et al., 2009). I have been examining the relationships between these different types of annotations (McCarthy, 2011).

Acknowledgments

I would like to thank all my collaborators on these projects. In alphabetical order these are John Carroll, Katrin Erk, Abhilash Inumella, Nick Gaylord, Spandana Gella, Ryu Iida, Bill Keller, Rob Koeling, Peng Jin, Aravind Joshi, Suresh Manadhar, Rada Mihalcea, Roberto Navigli, Mark Stevenson, Siva Reddy, Ravi Sinha, Sriram Venkatapathy, Julie Weeds and David Weir. I am indebted to Siva Reddy for a careful reading of an earlier draft. All the faults that remain are my own.

References

1. *Baldwin T., Bannard C., Tanaka T., Widdows D.* 2003. An Empirical Model of Multiword Expression Decomposability. Proceedings of the ACL Workshop on Multiword Expressions: Analysis, Acquisition and Treatment : 86–96
2. *Briscoe E., Carroll J.* 2002. Robust Accurate Statistical Annotation of General Text. Proceedings of the Third International Conference on Language Resources and Evaluation (LREC). : 1499–1504
3. *Church K.,Hanks P.* 1991. Word Association Norms, Mutual Information and Lexicography. Computational Linguistics, 16 (1): 22–29
4. *Dunning T.* 1993. Accurate Methods for the Statistics of Surprise and Coincidence. Computational Linguistics, 19 (1): 61–74
5. *Erk K.* 2007. A Simple, Similarity-based Model for Selectional Preferences. Proceedings of ACL 2007.
6. *Erk K., McCarthy D.* 2009. Graded Word Sense Assignment. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2009).
7. *Erk K., McCarthy D., Gaylord N.* 2009. Investigations on Word Senses and Word Usages. Proceedings of the Joint conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing ACL-IJCNLP.
8. *Erk K., Pado S.* 2010. Exemplar-Based Models for Word Meaning In Context.Proceedings of ACL 2010.
9. *Fellbaum C.* (editor).1998. WordNet, An Electronic Lexical Database.
10. *Gazdar G.* 1996. Paradigm merger in Natural Language Processing. Computing Tomorrow: Future Research Directions in Computer Science : 88–109
11. *Gale W., Church K., Iarovski D.* 1992. One Sense Per Discourse. Proceedings of the 4th DARPA Speech and Natural Language Workshop : 233–237
12. *Iida R., McCarthy D. and Koeling R.* 2008. Gloss-Based Semantic Similarity Metrics for Predominant Sense Acquisition. Proceedings of the Third International Joint Conference on Natural Language Processing : 561–568
13. *Jiang J., Conrath D.* Semantic similarity Based on Corpus Statistics and Lexical Taxonomy. 10th International Conference on Research in Computational Linguistics : 19–33
14. *Jin P., McCarthy D., Koeling R., Carroll J.* 2009. Estimating and Exploiting the Entropy of Sense Distributions. Proceedings of the North American Chapter of the Association for Computational Linguistics — Human Language Technologies (NAACL HLT) 2009 Conference
15. *Klavans J., Reznik P.* (editors.). 1996. The Balancing Act: Combining Symbolic and Statistical Approaches to Language.
16. *Koeling R., McCarthy D.* 2007. Sussx: WSD using Automatically Acquired Predominant Senses. Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007) : 314–317

17. *Koeling R., McCarthy D.* 2008. From Predicting Predominant Senses to Using Local Context for Word Sense Disambiguation. Semantics in Text Processing. STEP 2008 Conference Proceedings :129–138
18. *Koeling R., McCarthy D., Carroll J.* 2005. Domain-Specific Sense Distributions and Predominant Sense Acquisition. Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing : 419–426.
19. *Koeling R., McCarthy D. , Carroll J.* 2007. Text Categorization for Improved Priors of Word Meaning. Proceedings of the Eighth International Conference on Intelligent Text Processing and Computational Linguistics (CICLING 2007)
20. *Leech G.* 1992. 100 million Words of English: the British National Corpus. Language Research, 28(1):1–13
21. *Lesk M.* 1986. Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone From an Ice Cream Cone. Proceedings of the ACM SIGDOC Conference : 24–26
22. *Li H., Abe N.* 1998. Generalizing Case Frames Using a Thesaurus and the MDL Principle. Computational Linguistics, 24(2) : 217–244
23. *Lin D.* 1998. An Information-Theoretic Definition of Similarity. Proceedings of the 15th International Conference on Machine Learning.
24. *Lin D.* 1999. Automatic Identication of Noncompositional Phrases. Proceedings of ACL-1999 : 317–324.
25. *Mihalcea R., Sinha R., McCarthy D.* 2010. SemEval-2010 Task 2: Cross-Lingual Lexical Substitution. Proceedings of SemEval-2010: 5th International Workshop on Semantic Evaluations ACL 2010
26. *McCarthy D.* 2000. Using Semantic Preferences to Identify Verbal Participation in Role Switching Alternations. Proceedings of the first Conference of the North American Chapter of the Association for Computational Linguistics.
27. *McCarthy D.* 2011. Measuring Similarity of Word Meaning in Context with Lexical Substitutes and Translations. Computational Linguistics and Intelligent Text Processing 12th International Conference, CICLing 2011 : 238–252.
28. *McCarthy D., Keller B., Carroll J.* 2003. Detecting a Continuum of Compositionality in Phrasal Verbs. Proceedings of the ACL-SIGLEX Workshop on Multiword Expressions: Analysis, Acquisition and Treatment.
29. *McCarthy D., Keller, B., Navigli, R.* 2010. Getting Synonym Candidates from Raw Data in the English Lexical Substitution Task. Proceedings of the 14th EURALEX International Congress. Leeuwarden.
30. *McCarthy D., Koeling R., Weeds J., Carroll J.* 2004. Finding Predominant Senses in Untagged Text. Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics: 280–287.
31. *McCarthy D., Koeling R., Weeds J. ,Carroll J.* Unsupervised Acquisition of Predominant Word Senses. Computational Linguistics, 33 (4) : 553–590.
32. *McCarthy D., Navigli R.* 2009. The English Lexical Substitution Task. Language Resources and Evaluation 43 (2) Special Issue on Computational Semantic Analysis of Language: SemEval-2007 and Beyond : 139–159.

33. *McCarthy D., Venkatapathy S., Joshi A. K.* 2007. Detecting Compositionality of VerbObject Combinations using Selectional Preferences. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2007) : 369–379.
34. *Miller G. A., Leacock C., Tengi R., Bunker R. T.* 1993. A Semantic Concordance. Proceedings of the ARPA Workshop on Human Language Technology : 303–308.
35. *Mitchell J. , Lapata M.* 2008. Vector-based Models of Semantic Composition. Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies : 236–244.
36. *Navigli R.* 2009. Word Sense Disambiguation: a Survey. ACM Computing Surveys, 41(2) : 1–69.
37. *Ng H. T.* 1997. Getting Serious about Word Sense Disambiguation. Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How? :1–7.
38. *Pantel P., Lin D.* 2002. Discovering Word Senses from Text. Proceedings of ACM Conference on Knowledge Discovery and Data Mining (KDD-02) : 613–619.
39. *Patwardhan S., Pedersen T.* 2003. The CPAN WordNet::Similarity Package.// http://search.cpan.org/˜sid/WordNet-Similarity-0.05/ 2003
40. *Reddy S., Inumella I., McCarthy D., Stevenson M.* 2010. IIITH: Domain Specific Word Sense Disambiguation. Proceedings of SemEval-2010: 5th International Workshop on Semantic Evaluations.
41. *Reddy S., McCarthy D., Manandhar S., Gella, S.* 2011. Exemplar-based Word-Space Model for Compositionality Detection.
42. *Reznik P.* 1993. Selection and Information: A Class-Based Approach to Lexical Relationships.
43. *Rose T. G., Stevenson M., Whitehead M.* 2002. The Reuters Corpus Volume 1 — From Yesterday's News to Tomorrow's Language Resources. Proceedings of the Third International Conference on Language Resources and Evaluation : 827–833.
44. *Sag I., Baldwin T., Bond F., Copestake A., Flickinger D.* 2002. Multiword Expressions: A Pain in the Neck for NL. Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics : 1–15.
45. *Venkatapathy S., Joshi A. K.* 2005. Measuring the Relative Compositionality of Verb-Noun (V-N) Collocations by Integrating Features. Proceedings of the joint conference on Human Language Technology and Empirical methods in Natural Language Processing : 899–906.
46. *Weeds J.* 2003. Measures and Applications of Lexical Distributional Similarity.
47. *Weeds J., Weir D., McCarthy D.* 2004. Characterising Measures of Lexical Distributional Similarity. Proceedings of the 20th International Conference of Computational Linguistics : 1015–1021.