CONTENTS

Detecting Dangerous Coordination Ambiguities Using Word Distribution

Francis Chantree*, Alistair Willis*, Adam Kilgarriff** &
Anne de Roeck*

**The Open University, ***Lexical Computing Ltd*

Abstract

In this paper we present heuristics for resolving coordination ambiguities. We test the hypothesis that the most likely reading of a coordination can be predicted using word distribution information from a generic corpus. Our heuristics are based upon the relative frequency of the coordination in the corpus, the distributional similarity of the coordinated words, and the collocation frequency between the coordinated words and their modifiers. These heuristics have varying but useful predictive power. They also take into account our view that many ambiguities cannot be effectively disambiguated, since human perceptions vary widely.

1 Introduction

Coordination ambiguity is a very common form of structural (i.e., syntactic) ambiguity in English. However, although coordinations are known to be a "pernicious source of structural ambiguity in English" (Resnik 1999), they have received little attention in the literature compared with other structural ambiguities such as prepositional phrase (PP) attachment.

Words and phrases of all types can be coordinated (Okumura & Muraki 1994), with the external modifier being a word or phrase of almost any type and appearing either before or after the coordination. So for the phrase:

Assumptions and dependencies that are of importance

the external modifier *that are of importance* may apply either to both *assumptions* and *dependencies* or to just the *dependencies.*

We address the problem of disambiguating coordinations, that is, determining how the external modifier applies to the coordinated words or phrases (known as 'conjuncts'). We describe a novel disambiguation method using several types of word distribution information, and empirically validate this method using a corpus of ambiguous phrases, for which preferred readings were selected by multiple human judges. We also introduce the concept of an *ambiguity threshold* to recognise that the meaning of some ambiguous phrases cannot be judged reliably. All the heuristics use information generated by the Sketch Engine (Kilgarriff et al. 2004) operating on the British National Corpus (BNC) (`http://www.natcorp.ox.ac.uk`).

Throughout this paper, the examples have been taken from requirements engineering documents. Gause and Weinberg (1989) recognise requirements as a domain in which misunderstood ambiguities may lead to serious and potentially costly problems.

2 Methodology

'Central coordinators', such as *and* and *or*, are the most common cause of coordination ambiguity, and account for approximately 3% of the words in the BNC. We investigate single coordination constructions using these (and *and/or*) and incorporating two conjuncts and a modifier, as in the phrase:

old boots and shoes,

where *old* is the modifier and *boots* and *shoes* are the two conjuncts. We describe the case where *old* applies to both *boots* and *shoes* as 'coordination-first', and the case where *old* applies only to *boots* as 'coordination last'.

We investigate the hypothesis that the preferred reading of a coordination can be predicted by using three heuristics based upon word distributions in a general corpus. The first we call the *Coordination-Matches* heuristic, which predicts a coordination-first reading if the two conjuncts are frequently coordinated. The second we call the *Distributional-Similarity* heuristic, which predicts a coordination-first reading if the two conjuncts have strong 'distributional similarity'. The third we call the *Collocation-Frequency* heuristic, which predicts a coordination-last reading if the modifier is collocated with the first conjunct more often than with the second. We represent the conjuncts by their head words in all these three types of analysis.

In our example, we find that *shoes* is coordinated with *boots* relatively frequently in the corpus. *boots* and *shoes* are shown to have strong distributional similarity, suggesting that *boots and shoes* is a syntactic unit. Both these factors predict a coordination-first reading. Thirdly, the 'collocation frequency' of *old* and *boots* is not significantly greater than that of *old* and *shoes* and so a coordination-last reading is not predicted. Therefore, all the heuristics predict a coordination-first reading for this phrase.

In order to test this hypothesis, we require a set of sentences and phrases containing coordination ambiguities, and a judgement of the preferred reading of the coordinations. The success of the heuristics is measured by how accurately they are able to replicate human judgements. We obtained the sentences and phrases from a corpus of requirements documents, manually identifying those that contain potentially ambiguous coordinating conjunctions. Table 1 lists the sentences by part of speech of the head word of the conjuncts; Table 2 lists them by part of speech of the external modifier.

Head Word	% of Total	Example from Surveys (head words underlined)
Noun	85.5	Communication and performance requirements
Verb	13.8	Proceed to enter and verify the data
Adjective	0.7	It is very common and ubiquitous

Table 1: *Breakdown of sentences in dataset by head word type*

Modifier	% of Total	Example from Surveys (modifiers underlined)
Noun	46.4	(It) targeted the project and election managers
Adjective	23.2	.... define architectural components and connectors
Prep	15.9	Facilitate the scheduling and performing of works
Verb	5.8	capacity and network resources required
Adverb	4.4	(It) might be automatically rejected or flagged
Rel. Clause	2.2	Assumptions and dependencies that are of importance
Number	0.7	zero mean values and standard deviation
Other	1.4	increased by the lack of funding and local resources

Table 2: *Breakdown of sentences in dataset by modifier type*

Ambiguity is context-, speaker- and listener-dependent, so there are no absolute criteria for judging it. Therefore, rather than rely upon the judgement of a single human reader, we took a consensus from multiple readers. This approach is known to be very effective albeit expensive (Berry 2003).

In total, we extracted 138 suitable coordination constructions and showed each one to 17 judges. They were asked to judge whether each coordination was to be read coordination first, coordination last or "ambiguous so that it might lead to misunderstanding". In the last case, the coordination is then classed as an '*acknowledged ambiguity*' for that judge. We believe that by using a sufficiently large number of judges, we can estimate how certain we can be that the coordination should be read in a particular way. Then we use the idea of an adjustable 'ambiguity threshold', which represents the minimum acceptable level of certainty about the preferred reading of a passage of text in order for it not to be considered ambiguous.

3 Related research

There is little work on automatically disambiguating coordination ambiguities in English. What research there has been addresses several different tasks, illustrating the difficulty of a full treatment of all ambiguities caused by coordinations. For instance, Agarwal and Boggess (1992) developed a method of recognising which phrases are conjoined by matching part of speech and case labels in a tagged dataset. They achieved an accuracy of 82.3% using the machine-readable Merck Veterinary Manual as their dataset. In a full system, their methods would form a useful initial step for identifying the coordinated structures, before attempting to determine attachment. Goldberg (1999) adapted Ratnaparkhi's (1998) PP attachment method for use on coordination ambiguities. She achieved an accuracy of 72% on the annotated attachments of her test set, drawn from the Wall Street Journal by extracting head words from chunked text.

Resnik (1999) investigated the role of semantic similarity in resolving nominal compounds in coordination ambiguities of the form *noun1 and noun2 noun3*, such as *bank and warehouse guard*. To disambiguate, Resnick compares the relative information content of the classes in WordNet that subsume the noun pairs; this method has achieved 71.2% precision and 66.0% recall of the correct human disambiguations in a dataset drawn from the Wall Street Journal. By adding an evaluation of the selectional association between the nouns to his semantic similarity evaluation, Resnick achieves precision of 77.4% and 69.7% recall on complex coordinations of the form *noun0 noun1 and noun2 noun3*.

We believe that because our method is applicable to any part of speech for which word distribution information is available, our results are more generally applicable than those of Resnick, which are applied specifically to nominal compounds. In addition, we do not know of other comparable work in which multiple readers have been used to select a preferred reading. This approach to collecting our datasets gives us an additional insight into the relative certainty of different readings.

4 Disambiguation empirical study

We maximise our heuristics' performance using ambiguity thresholds and ranking cut-offs. The ambiguity threshold is the minimum level of certainty that must be reflected by the consensus of survey judgements. Suppose a coordination is judged to be coordination-first by 65% of judges, and we use a heuristic that predicts coordination-first readings. Then, if the ambiguity threshold is 60% the consensus judgement will be considered to be coordination-first, whereas it will not if the ambiguity threshold is 70%. This can significantly change the baseline — the percentage of either coordination-first or coordination-last judgements, depending on which of these readings the heuristic is predicting. The ranking cut-off is the point below which a heuristic is considered to give a negative result. We use data in the form of rankings as these are considered more accurate than frequency or similarity scores for word distribution comparisons (McLaughlan 2004).

True positives for a heuristic are those coordinations for which it predicts the consensus judgement. Precision for a heuristic is the number of true positives divided by the number of positive results it produces; recall is the number of true positives divided by the number of coordinations it should have judged positively. Precision is much more important to us than recall: we wish each heuristic to be a reliable indicator of how a coordination should be read, and hope to achieve good recall by the heuristics having complementary coverage. We use a weighted f-measure statistic (van Rijsbergen 1979) to combine precision and recall — with $\beta = 0.25$, strongly favouring precision — and seek to maximise this for all of our heuristics:

$$F\text{–}Measure = \frac{(1+\beta) * Precision * Recall}{\beta^2 * Precision + Recall}$$

We employ 10-fold ‘cross validation’, to avoid the problem of ‘overfitting’ (Weiss & Kulikowski 1991). Our dataset is split into ten equal parts, nine of which are used for training to find the optimum ranking cut-off and ambiguity threshold for each heuristic. (The former are found to be the same for all 10 folds for all three heuristics.) The heuristics are then run on the heldout tenth part using those cut-offs and ambiguity thresholds. This procedure is carried out for each heldout part, and the heuristics’ performances over all the iterations are averaged to give their overall performances.

4.1 *Our tools*

All our heuristics use statistical information generated by the Sketch Engine with the BNC as its data source. The BNC is a modern corpus of over 100 million words of English, collated from a variety of sources. The Sketch Engine provides a thesaurus giving distributional similarity between words, and word sketches giving the frequencies of word collocations in many types of syntactic relationship. It accepts input of verbs, nouns and adjectives. In the word sketches, head words of conjuncts are found efficiently by using grammatical patterns (Kilgarriff et al. 2004).

The Sketch Engine’s thesaurus is in the tradition of Grefenstette (1994); it measures distributional similarity between any pair of words according to the number of corpus contexts they share. Contexts are shared where the relation and one collocate remain the same, so ⟨*object, drink, wine*⟩ and ⟨*object, drink, beer*⟩ count towards the similarity between *wine* and *beer*. Shared collocates are weighted according to the product of their mutual information, and the similarity score is the sum of these weights across all shared collocates, as in (Lin 1998). Distributional thesauruses are well suited to our task, as words used in similar contexts but having dissimilar semantic meaning, such as *good* and *bad*, are often coordinated.

4.2 *Coordination-matches heuristic*

We hypothesise that if a coordination is found frequently within a corpus then a coordination-first reading is the more likely. We search the BNC for each coordination in our dataset using the Sketch Engine, which provides lists of words that are conjoined with *and* or *or*. Each head word is looked up in turn. The ranking of the match of the second head word with the first head word may not be the same as the ranking of the match of the first head word with the second head word. This is due to differences in the overall frequencies of the two words. We use the higher of the two rankings. We find that considering only the top 25 rankings is a suitable cut-off. An ambiguity threshold of 60% is found to be the optimum for all ten folds in the cross-validation exercise. For the example from our dataset:

Security and Privacy Requirements,

the higher of the two rankings of *Security* and *Privacy* is 9. This is in the top 25 rankings so the heuristic yields a positive result. The survey judgements were: 12 coordination-first, 1 coordination-last and 4 ambiguous, giving a certainty of $12/17 = 70.5\%$. As this is over the ambiguity threshold of 60%, the heuristic always yields a true positive result on this sentence.

Averaging over all ten folds, this heuristic achieves 43.6% precision, 64.3% recall and 44.0% f-measure. However, the baselines are low, given the relatively high ambiguity threshold, giving 20.0 precision and 19.4 f-measure percentage points above the baselines.

4.3 *Distributional-similarity heuristic*

Our second hypothesis follows a suggestion by Kilgarriff (2003) that if two conjuncts display strong distributional similarity, then the conjunction is likely to form a syntactic unit, giving a coordination-first reading.

For each coordination, the lemmatised head words of both the conjuncts are looked up in the Sketch Engine's thesaurus. We use the higher of the ranking of the match of the second head word with the first head word and the ranking of the match of the first head word with the second head word. The optimal cut-off is to consider only the top 10 matches. An ambiguity threshold of 50% produces optimal results for 7 of the folds, while 70% is optimal for the other 3. For the example from our dataset:

processed and stored in database,

the verb *process* has the verb *store* as its second ranked match in the thesaurus, and vice versa. As this is in the top 10 matches, the heuristic yields a positive result. The survey judgements were: 1 coordination-first, coordination-last and 5 ambiguous, giving a certainty of $1/17 = 5.9\%$. As this is below both the ambiguity thresholds used by the folds, the heuristic's performance on this sentence always yields a false positive result.

Averaging for all ten folds, this heuristic achieves 50.8% precision, 22.4% recall and 46.4% f-measure, and 11.5 precision and 5.8 f-measure percentage points above the baselines.

Heuristic	Re-	Baseline	Prec.	Prec.	F-meas.	F-meas.
	call	Precision		above base	($\beta = 0.25$)	above base
(1) Coordination-match	64.3	23.6	43.6	20.0	44.0	19.4
(2) Distrib-similarity	22.4	39.3	50.8	11.5	46.4	5.8
(3) Collocation-freq.	35.3	22.1	40.0	17.9	37.3	14.1
(4) = (1) & not (3)	64.3	23.6	47.1	23.5	47.4	22.9

Table 3: *Performance of our heuristics (%)*

4.4 *Collocation-frequency heuristic*

Our third heuristic predicts coordination-last readings. We hypothesise that if a modifier is collocated in a corpus much more frequently with the conjunct head word that it is nearest to than it is to the further head word, then it is more likely to form a syntactic unit with only the nearest head word. This implies that a coordination-last reading is the more likely.

We use the Sketch Engine to find how often the modifier in each sentence is collocated with the conjuncts, head words. We experimented with collocation ratios, but found the optimal cut-off to be when there are no collocations between the modifier and the further head word, and any non-zero number of collocations between the modifier and the nearest head word. An ambiguity threshold of 40% produces optimum results for 8 of the folds, while 70% is optimal for the other 2. For the example from our dataset:

project manager and designer,

project often modifies *manager* in the BNC but never *designer*, and so the heuristic yields a positive result. The survey judgements were: 8 coordination-last, 4 coordination-first and 5 ambiguous, giving a certainty of 8/17 = 47.1%. This is over the ambiguity threshold of 40% but under the threshold of 70%. On this sentence, the heuristic therefore yields a true positive result for 8 of the folds but a false positive result for 2 of them.

Averaging for all ten folds, the heuristic achieves 40.0% precision, 35.3% recall and 37.3% f-measure, and 17.9 precision and 14.1 f-measure percentage points above the baselines.

5 Evaluation and discussion

Table 3 summarises our results. Our use of ambiguity thresholds prevents readings being assigned to highly ambiguous coordinations. This has two contrary effects on performance: the task is made easier as the target set contains more clear-cut examples, but harder as there are fewer examples to find. Our precision and f-measure in terms of percentage points over the baselines, except for the distributional-similarity heuristic, are encouraging.

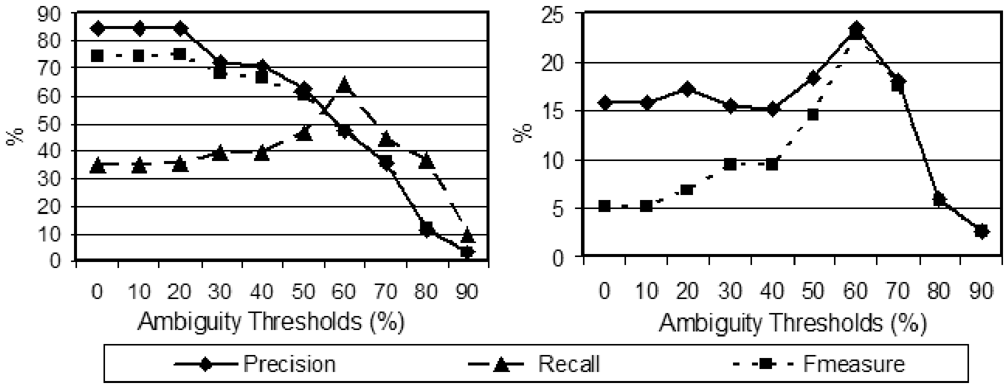

Fig. 1: *Heuristic 4: Left graph – absolute performance; Right graph – performance as percentage points over baselines*

We combine the two most successful heuristics, shown in the last line of Table 3, by saying a coordination-first reading is predicted if the coordination-matches heuristic gives a positive result and the collocation-frequency heuristic gives a negative one. The left hand graph of Figure 1 shows the precision, recall and f-measure for this fourth heuristic, at different ambiguity thresholds. As can be seen, high precision and f-measure can be achieved with low ambiguity thresholds, but at these thresholds even highly ambiguous coordinations are judged to be either coordination-first or -last. The right hand graph of Figure 1 shows performance as percentage points above the baselines. Here the fourth heuristic performs best, and is more appropriately used, when the ambiguity threshold is set at 60%.

Instead of using the optimal ambiguity threshold, users of our technique can choose whatever threshold they consider appropriate, considering how critical they believe ambiguity to be in their work. Figure 2 shows the proportions of ambiguous and non-ambiguous interpretations at different ambiguity thresholds. None of the coordinations are judged to be ambiguous with an ambiguity threshold of zero — which is a dangerous situation — whereas at an ambiguity threshold of 90% almost everything is considered ambiguous.

6 Conclusions and further work

Our results show that the collocation-frequency heuristic and (particularly) the coordination-matches heuristic are good predictors of the preferred reading of a sentence displaying coordination ambiguity, and that combining them increases performance further. However, the performance of

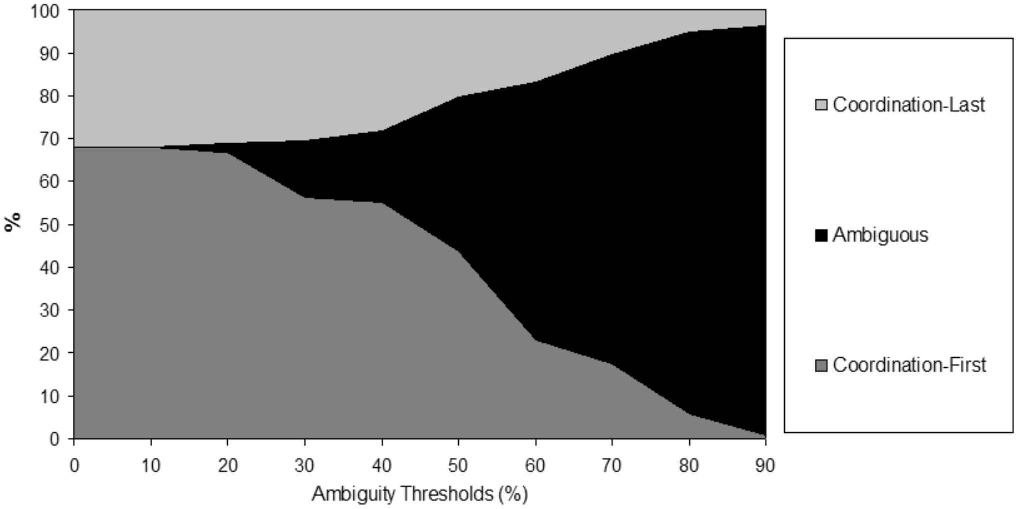

Fig. 2: *Ambiguous and non-ambiguous readings at different thresholds*

the distributional-similarity heuristic suggests that distributional similarity between head words of conjuncts is only a weak indicator of preferred readings.

The success of these heuristics is perhaps surprising, as the distribution information was obtained from a general corpus (the BNC), but tested on a specialist data set (requirements documents). This indicates that many distributions of head words in the data set are reflected in the corpus. These are promising results, as they suggest that our techniques may be applicable across different domains of discourse, without the need for distribution information for specialist corpora. The results also show that the heuristics are not specific to grammatical constructions: the method is applicable to coordinations of different types of word, and different types of modifier.

We have found that people's judgements can vary quite widely: different people interpret a sentence differently, but do not themselves consider the sentence ambiguous. We call this '*unacknowledged ambiguity*'; it is potentially more dangerous than acknowledged ambiguity as it is not noticed and therefore may not be resolved. Unacknowledged ambiguity is measured as the number of judgements in favour of the minority non-ambiguous choice, over all the non-ambiguous judgements. The average unacknowledged ambiguity over all the examples in our dataset is 15.3%.

This paper is part of wider research into notifying users of ambiguities in text and informing them of how likely they are to be misunderstood. We are currently testing heuristics based on morphology, typography and word sub-categorisation. In this work we investigate the multi-level conjunct parallelism model of Okumura and Muraki (1994).

REFERENCES

Agarwal, Rajeev & Lois Boggess. 1992. "A Simple but Useful Approach to Conjunct Identification". *Proceedings of the 30th Conference on Association for Computational Linguistics*, 15-21. Newark, Delaware.

Berry, Daniel & Erik Kamsties & Michael Krieger. 2003. *From Contract Drafting to Software Specification: Linguistic Sources of Ambiguity. A Handbook.* `http://se.uwaterloo.ca/ dberry/handbook/ambiguityHandbook.pdf`

Gause, Donald C. & Gerald M. Weinberg. 1989. *Exploring Requirements: Quality Before Design.* New York: Dorset House.

Goldberg, Miriam. 1999. "An Unsupervised Model for Statistically Determining Coordinate Phrase Attachment". *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, 610-614. College Park, Maryland.

Grefenstette, Gregory. 1994. *Explorations in Automatic Thesaurus Discovery.* Boston, Mass.: Kluwer Academic.

Kilgarriff, Adam. 2003. "Thesauruses for Natural Language Processing". *Proceedings of Natural Language Processing and Knowledge Engineering* (*NLP-KE*) ed. by Chengqing Zong. 5-13. Beijing, China.

Kilgarriff, Adam & Pavel Rychly & Pavel Smrz & David Tugwell. 2004. "The Sketch Engine". *11th European Association for Lexicography International Congress* (*EURALEX 2004*), 105-116. Lorient, France.

Lin, Dekang. 1998. "Automatic Retrieval and Clustering of Similar Words". *Proceedings of the 17th International Conference on Computational Linguistics*, 768-774. Montreal, Canada.

McLauchlan, Mark. 2004. "Thesauruses for Prepositional Phrase Attachment". *Proceedings of Eight Conference on Natural Language Learning* (*CoNLL*) ed. by Hwee Tou Ng & Ellen Riloff, 73-80. Boston, Mass.

Okumura, Akitoshi & Kazunori Muraki. 1994. "Symmetric Pattern Matching Analysis for English Coordinate Structures". *Proceedings of the 4th Conference on Applied Natural Language Processing*, 41-46. Stuttgart, Germany.

Ratnaparkhi, Adwait. 1998. "Unsupervised Statistical Models for Prepositional Phrase Attachment". *Proceedings of the 17th International Conference on Computational Linguistics*, 1079-1085. Montreal, Canada.

Resnik, Philip. 1999. "Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language". *Journal of Artificial Intelligence Research* 11:95-130.

van Rijsbergen, C. J. 1979. *Information Retrieval.* London, U.K.: Butterworths.

Weiss, Sholom M. & Casimir A. Kulikowski. 1991. *Computer Systems that Learn: Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning, and Expert Systems.* San Francisco, Calif.: Morgan Kaufmann.

Index of Subjects and Terms