The Contribution of Domain-independent Robust Pronominal Anaphora Resolution to Open-Domain Question-Answering

Rebecca Watson, Judita Preiss & Ted Briscoe
Computer Laboratory
University of Cambridge
JJ Thomson Ave
Cambridge
CB3 OFD, UK
firstname.lastname@cl.cam.ac.uk

Abstract

We explore the performance increase that results from the use of robust domain-independent pronoun resolution as a component of an open-domain question-answering (QA) system. We describe a baseline system based on robust parsing, named entity recognition, and matching of underspecified (rMRS) semantic structures between question and putative answer sentences, and its performance on the TREC8 QA task. We derive an experimental upper bound for improvement on this task through use of a 2 sentence context window around putative answer sentences. We describe our pronoun resolver, and its integration with the baseline system. Finally, we assess the potential and actual improvement in QA given the performance of anaphora resolution on this data and methods for integration with the QA system.

1 Introduction

Open-domain question-answering (QA) involves matching a representation obtained from a user's question against a document collection in order to find appropriate short answers. It is intuitively clear, at least for questions which do not contain many equivalent correct answers in the collection, that intelligent use of the context around sentences which match elements of the question should improve performance. To give one example from the TREC-8 QA dataset, the question, *What country is the biggest producer of tungsten?* can be given the answer *China* using the following passage:

> The 15 countries attending the three-day annual market review, which ended yesterday, account for about 90 per cent of world trade in tungsten products. They include China, the biggest producer, which represents over 60 per cent of world trade...

However this requires the inference that China is a country and the biggest producer *of tungsten* to achieve a convincing match with the question. The first inference can be assisted by resolving the pronominal anaphor *They* to its antecedent *The 15 countries* and the second via a more complex inference from *producer* back to *tungsten products*.

We focus on the ability of *pronominal* anaphora resolution technology to provide enriched representations of sentences containing target answers which can be used as the basis for the first type of inference. We describe a baseline QA system based on robust parsing, named entity recognition, and matching of underspecified (rMRS) semantic structures between question and putative answer sentences, and its performance on the TREC8 QA task. We derive an experimental upper bound for improvement on this task through use of a 2 sentence context window around putative answer sentences. We describe our pronoun resolver and its integration with the baseline system. Finally, we assess the likely improvement in QA given the performance of anaphora resolution and method of integration into the QA sys-

tem.

2 The Baseline QA System

Our baseline system utilses the LingoERG grammar and the LKB parser (Copestake and Flickinger, 2000) to analyse questions and construct compositional semantic representations of them using minimal recursion semantics (MRS) (Copestake et al., 1999). Robust MRS (rMRS) is a syntactic variant of MRS which supports factorization and systematic underspecification of most aspects of this semantic representation (Briscoe et al., 2002). rMRS representations are factored into a set of elementary predications (e.g. **tungsten(x1)**) and a set of (sorted) variable equality statements (e.g. **x1=x2**). Variables can be event or object arguments of predicates which can be underspecified in cases where, say, it is unclear whether a PP is adverbial or adjectival, or can be handles used to specify scope. For our experiments we used 163 of the TREC 8 questions which the LKB analysed and assigned rMRS representations.

The 1000 top-ranked documents for the TREC 8 QA task were parsed using the RASP (robust accurate statistical parsing) system (Briscoe and Carroll, 2002). The highest-ranked analysis returned by RASP for each sentence was converted to a rMRS. Many of these rMRS representations are severely underspecified because of the impoverished information about lexical semantic types yielded by this system. In addition, about 23% of grammatical relations recovered by the RASP system in highest-ranked analyses are incorrect on test data. Therefore, we can assume that a similar proportion of the elements of the rMRS recovered from the syntactic analysis are likely to be incorrect. For a small fraction of very long and complex sentences, lists, and so forth rMRSs are based entirely on the part-of-speech tag of each word as the parser timed out before returning any syntactic analysis. An example of the output of RASP is given in Figure 1. The resulting (simplified) rMRS is underspecified for the type of the (event) variable (u4), the object (ARG2) of *include*, and the coreference between *China* and the following appositive NP, etc. The baseline QA system is based on checking the graded compatibility of rMRS elements between questions

```
(|T/txt-sc1/---|
 (|T/leta_s|
  (|S/s_co_np1|
   (|S/np_vp| |They_PPHS2|
    (|V/np| |include_VV0|
     (|NP/n1_name/-|
        (|N1/n| |China_NP1|))))
   |,_,|
   (|NP/det_n| |the_AT|
    (|N1/ap_n1/-|
     (|AP/a1| (|A1/a| |biggest_JJT|))
     (|N1/n| |producer_NN1|))))
  (|Tacl/comma-e| |,_,|
   (|S/whnp_vp| |which_DDQ|
    (|V/np| |represent+s_VVZ|
     (|NP/ap2_np| (|A1/a| |over_RP|)
      (|NP/plu3|
       (|N1/num2_nms|
          (|NP/num| (|N1/n| 60_MC))
        (|N1/nms_nms| |per_NNU|
         (|N1/n_of| |cent_NNU|
          (|PP/p1|
           (|P1/p_n1| |of_IO|
            (|N1/n1_nm| |world_NN1|
               (|N1/n| |trade_NN1|
                  ))))))))))))
GRs:
(ncsubj represent+s_VVZ which_DDQ _)
(dobj represent+s_VVZ cent_NNU _)
(ncsubj include_VV0 They_PPHS2 _)
(dobj include_VV0 China_NP1 _)
(ncmod _ producer_NN1 biggest_JJT)
(detmod _ producer_NN1 the_AT)
(ncmod _ include_VV0 producer_NN1)
(ncmod _ trade_NN1 world_NN1)
(ncmod of_IO cent_NNU trade_NN1)
(ncmod _ cent_NNU per_NNU)
(ncmod _ cent_NNU 60_MC)
(mod _ cent_NNU over_RP)
(cmod _ include_VV0 represent+s_VVZ)

rMRS:
they_rel u2, include_rel u4
ARG1  u4 u2, ARG2  u4 u7
china_rel x6, the_rel x12
biggest_rel x12, producer_rel x12
which_rel x27, represent_rel e29
over_rel e29, 60_rel u33
per_rel x35, cent_rel x37
of_rel e39, ARG2 e39 x41
world_rel x41, trade_rel x50
```

Figure 1: System Outputs

rMRS	0.472
+Morph	0.476
+WordNet+NE	0.484
rMRS+Context	0.619

Table 1: Baseline System(s) MRR Performance

rMRS	0.150
+Morph	0.178
+WordNet+NE	0.270
+Context	0.470

Table 2: Performance on Unseen Sentences

and putative answer sentences. The TREC 8 QA dataset and evaluation system was used to optimize the weights and matching criteria which determine the overall level of compatibility when matching different components of the rMRS structure. This resulted in a system which ignored elementary predications for closed class items, abstracted over subtypes of major parts-of-speech, used morphological analysis to match morphologically-related predicates (e.g. **producer** and **production**), WordNet hyponym and synonym variants of predicates (e.g. **inhabitant** $\rightsquigarrow$ **soul**), and named entity (NE) recognition to filter out entire rMRSs which did not contain a NP of the same sort as the question. The novel element to this baseline QA system is that it utilizes two extant parsers to generate rMRS representations for questions and document sentences and uses rMRS as the prime representation level upon which all assessments of compatibility are made.

3 Performance on the TREC 8 QA Task

Since our baseline system was optimized on the TREC 8 data, its performance on this data is not a realistic assessment of its general utility. Table 1 gives the mean reciprocal rank (MRR) for the baseline rMRS matching system, the baseline system with morphological analysis, and the baseline system with WordNet and NE filtering. The final column gives the results for our simplest baseline system with the addition of two sentence contexts to either side of the matching sentence, where all 5 sentences are submitted to the evaluation system[1]. It is clear from these results that, whilst the use of morphological and semantic relationships (from WordNet) and semantic filtering via NE sorts does improve performance, a far more dramatic improvement is possible by inclusion of further context. Note that submitting the entire context to the evaluation code is a crude way of assessing an upper (experimentally-derived) bound on correct exploitation of this context in the 50byte task. Assuming 100% exploitation of context in this way reduces the number of unanswered questions by about 33%.

We have also tested various variants of our baseline system on a small sample of unseen data from the TREC 9 QA track and results for these 10 sentences do support the conclusion that the matching criteria and weights selected are valid – see Table 2. If we could effectively exploit context to derive the correct short answer, then this would improve the performance of the system significantly and this would produce a bigger increase in performance (potentially) than focussing on variants of predicates in matching.

4 The Utility to QA

Analysis of the top 5 matching contexts returned by the baseline system revealed that there are 1041 third person pronouns in the contexts (roughly 1.2% of the total number of words). This suggests that anaphora resolution might be able to enrich the representation of the highest-matching sentence so that the correct short answer could be directly extracted via this sentence. For instance, returning to the example in section 1, resolving *They* to *The 15 countries* would license the addition of an equality statement between the rMRS variables associated with these two NPs, and this would yield an enriched rMRS for the sentence containing *China* pointing to an elementary predication for **countries**. Accurate anaphora resolution will improve the match, but there remain further inferences to be made before we can guarantee that this sentence will become the highest-ranked match to the question rMRS.

We evaluated all the matching contexts, given the baseline rMRS system, to determine whether anaphora resolution would potentially improve performance by increasing the compatibility of ques-

[1]Morton (2000) reports that a look-back of 2 sentences makes the antecedent of a pronoun accessible 98.7% of the time.

intraP	0.11
interP	0.04
interD	0.13
contx+	0.14
contx-	0.10

Table 3: Proportions of Contexts Improvable

tion rMRS and the enriched rMRS of the sentence containing the answer in the context. Table 3 classifies the proportion of contexts in which resolution of intrasential pronominal anaphora (intraP), intersentential pronominal anaphora (interP), and intersentential definite description (interD) anaphora could improve QA performance in this sense.[2] It also indicates where more open-ended contextual inference could, in principle, work (contx+) and where the context is not sufficient (contx-, i.e. contexts where the match to the answer is effectively spurious). The remaining proportion (0.48) of contexts not classified in Table 3 contained correct answers consisting of phrases, often appositives NPs, in the matching target sentence which also included all the information required to match the question and generate the correct short answer. This high proportion may reflect the artificial method in which the TREC 8 QA track questions were created. However, if the remaining data are representative, they suggest that anaphora resolution has a significant role to play in the exploitation of context. Ignoring the 10% of spurious contexts, anaphora resolution is potentially beneficial in two thirds of the remaining contextual cases, and about 30% of the TREC 8 questions overall. This result does not imply that perfect anaphora resolution would improve performance by one third or better – only a small number of sentences containing correct short answers would be sufficently enriched by resolving anaphors to support straightforward generation of the short answer by rMRS matching. Nevertheless, it does suggest that anaphora resolution is a necessary and potentially significant component for the effective exploitation of context in a QA system.

[2]We are assuming that cases of intrasentential definite description anaphora, such as in appositives, will be handled correctly by the parser and rMRS pattern matcher.

5 Robust Domain-Independent Anaphora Resolution

Accurate anaphora resolution requires access to syntactic information in order to compute non-coreference constraints and to compute the structural salience of potentially coreferential antecedents. Work by Lappin and McCord (1990) and Lappin and Leass (1994) demonstrated that reasonable performance could be achieved, given an accurate and detailed enough syntactic analysis from which predicate argument structure (or deep grammatical relations including 'understood' control relations) could be extracted. Kennedy and Boguraev (1996) demonstrated that robust anaphora resolution could be achieved, using tags encoding lexical syntactic category and surface grammatical relations to compute salience but not non-coreference constraints. Castano et al. (2002) demonstrated that an even more syntactically impovershed system only utilizing part-of-speech tags could achieve reasonable performance, particularly on anaphoric definite descriptions, if supplemented with rich domain information in the form of sortal constraints. Ge et al. (1998) developed a probabilistic approach which integrated both types of syntactic factor with many others, such as distance and mention rate of the antecedent, as well as domain-dependent lexical information. All these systems achieved accuracy rates of 70% or better. However, meaningful comparison is hard because of the slightly different evaluation schemes employed and very different datasets used.

Preiss and Briscoe (2003) describe a reimplementation of the system of Lappin and Leass (1994) (hereafter LL) using the grammatical relations (GR) output from the RASP system. The RASP system makes very little use of lexical information (unlike most statistical parsers) in an attempt to remain (initially) as domain-independent as possible. As the system returns ranked analyses, this does not preclude reranking utilizing domain-dependent (lexical) information where this is available, but our goal is to achieve useful levels of accuracy without relying on the availability of domain-specific lexical resources. We have pursued a similar approach to anaphora resolution, attempting to exploit structural information fully in order to produce as ac-

Factor	Weight
Sentence recency	100
Subject emphasis	80
Existential emphasis	70
Accusative emphasis	50
Indirect object/oblique	40
Head noun emphasis	80
Non-adverbial emphasis	50
Parallelism	35
Cataphora	175

Table 4: LL Salience Weights

curate a domain-independent ranking of potential antecedents as possible. As yet, our implementation does not cover definite descriptions, although it would be possible, in principle, to do so by utilizing WordNet or another relatively domain-independent lexical resource to capture the synonymy and hyponymy relations required to evaluate sortal compatibility between antecedents and anaphoric definite descriptions.

The RASP GR scheme utilizes 20 GRs organized into an inheritance network (see Carroll et al. (2003)). An example of GR output is given in Figure 1 above. LL's non-coreference constraints can be captured effectively in terms of this GR scheme. For example, RASP output for *Kim seems to want to see him* includes the GRs:

```
(ncsubj see_VV0 Kim_NP1 _)
(dobj see_VV0 he_PPHO1 _)
```

and LL's argument domain filter can be succinctly encoded as:

```
(arg - X N -)
(arg - X P -)
```

where arg $\in \{ncsubj, dobj, iobj, obj2\}$, X is a variable over predicates, and N and P are nominal and pronominal dependents of X respectively. Thus *Kim* and *him* are predicted to be non-coreferential because they are both arguments (dependents) of the same (head) predicate *see*. All of LL's non-coreference and agreement filters can be implemented in terms of such patterns over RASP GR output (see Preiss and Briscoe (2003)).

LL's salience factors and weights are given in Table 4. The factors are straightforwardly computed from RASP GR output and the overall salience of a potential antecedent is a weighted sum of these factors. The original LL weights were manually set on the basis of heuristic experiments with their data.

	BC	BU	CH	C1	C2
1	60	63	63	63	61
2	51	53	54	55	54
3	70	70	69	67	69
4	67	65	70	64	67
5	55	53	50	52	52
μ	61	61	62	61	61

Table 5: Anaphora Results

Preiss and Briscoe (2003) and Preiss (2002) report experiments on a new annotated corpus, drawn from the BNC (Leech, 1992), containing 2400 manually-resolved pronominal anaphors. The corpus was divided into 5 sections containing roughly equal numbers of anaphors. GR output was obtained for the anaphora corpus from five statistical parsers: RASP (BC), Buchholz (2002) (BU), Charniak (2000) (CH), Collins (1997) model 1 (C1) and model 2 (C2). For the latter three parsers, we implemented GR extraction rules for Penn Treebank analyses. In Table 5, we present the results (precision only as all pronouns are always attempted, mean μ) obtained from our implementation of the LL algorithm using the output from all five parsers. These results revealed no significant parser differences in performance on anaphora resolution. However, given that the RASP system is the least lexicalized of the five evaluated, this result suggests that state-of-the-art and domain-independent anaphora resolution can be achieved this way. Error analysis does, however, reveal that the LL algorithm is not optimal. For example, the argument-contained filter, used to prevent *He* and *the man* coreferring in examples like *He believes that the man is amusing*, removed the correct antecedent in 12 sentences with intrasentential anaphors, despite the correctness of the GR output, and similarly 8 antecedents were ruled out by hard agreement constraints.

6 Performance on the QA Contexts

In order to evaluate the utility of incorporating the RASP-GR+LL algorithm described above into

our baseline QA system, we parsed the contexts identified in Table 3 which contained pronominal anaphors and ran the LL algorithm on the GR output. Manual evaluation of the output revealed that 73.2% of these anaphors were correctly resolved by the system. Error analysis revealed that 36% of the errors were caused by misidentification of the head of the antecedent rather than misidentification of the antecedent (e.g. *El* instead of *El Nino*). Several errors were a consequence of chain effects where an initial anaphor was incorrectly resolved to a full NP antecedent and subsequent anaphors, though correctly resolved to the initial anaphor, then 'inherited' the incorrect full antecedent through this chain[3].

7 The Contribution to QA

To assess the contribution to performance that a domain-independent anaphora resolution component would make to a QA system, we augmented the rMRS provided to our baseline QA system by adding variable equality statements binding antecedent and anaphor, as found by the resolution component, in the rMRS representations, which were then fed to the compatibility matching component of the QA system.

In the cases where the correct antecedent is found the potential effects are two-fold. Firstly, the target sentence may receive a higher ranking because of the higher degree of resultant compatibility between question rMRS and that of the target sentence. Secondly, elementary predications obtained via the antecedent may enable the QA system to directly construct an appropriate short answer from the target sentence. Conversely, ranking may decline if antecedents are incorrect, or if variable linking results in irrelevant rMRSs being added.

We assessed the effect on ranking by providing several enriched rMRS representations to the compatibility matching component (along with all the rMRSs for the rest of the top-ranked documents). We assessed the ability of the system to extract a short answer from the highest-ranked sentences by submitting these sentences together with the textual forms associated with the additional rMRSs to the evaluation system.

[3]We have not directly tested the alternative architecture in which anaphora resolution is applied to the entire document collection (as with parsing) prior to any matching. Our assumption is that, given the level of performance of the anaphora resolution component, this would result in additional noise and degraded performance over its focussed application in shorter matching contexts.

Baseline	0.491
+antecedent	0.510
+direct-subst	0.499
+partial-rMRS	0.483
+full-rMRS	0.459
+context	0.619

Table 6: MRR with Anaphora

In Table 6 row 'rMRS' gives the MRR score for a slightly optimized baseline system (as described in section 3) and rMRS+context restates the experimental upper bound achievable through optimal exploitation of the context window (with the caveat that it is an overestimate due to spurious answer matching, see section 4). Row +antecedent reports the MRR score obtained by manually substituting the antecedent found by the anaphora resolution system for the pronoun(s) in the text of the target submission sentence. This provides an upper bound on performance increase, modelling optimal integration of information from the antecedent constituent. Row +direct-subst reports the MRR obtained by automatically enriching the rMRS for the target sentences with the elementary predications corresponding to the head (nouns) of antecedents of anaphors in the context window. Row +partial-rMRS gives results when the rMRS for the target sentence is automatically enriched not only with the rMRS corresponding to the antecedent head but also any other elementary predications and argument constraints on the variable introduced by the antecedent head elementary predication. Row +full-rMRS gives the MRR for a system which integrates the entire rMRS for a context sentence containing an antecedent, and that of its preceeding sentence if it is also anaphorically linked to it or the target sentence.

The improvement in MRR for manual integration of the antecedent illustrates that pronoun resolution has the potential to improve ranking (and identification) of potential answers that would otherwise not be detected. Returning to our previous example, the baseline QA system found the optimal sentence:

Tungsten producing and consuming coun-

tries have been meeting this week in Geneva...

The approach correctly resolved the pronoun *They* in the correct sentence to *country* in the passage. Integrating these rMRS structures increases the ranking (match score) for the correct sentence: *They* (country) *include China, the biggest producer...* above that of the previous (incorrect) sentence. Further analysis demonstrated that 71% of the submissions are improved by this method. In 90% of the cases, this improvement was not reflected in terms of MRR score as the anaphora resolution occurred for sentences in which the resolution referred to an antecedent from the same (correct) submission sentence. For QA in general, however, this result is still important as the improved context can be used during NE recognition and short answer construction. This example also highlights that integration of further information is potentially beneficial. The ranking of the correct sentence would further improve if the integration cintext and target sentences also included the 'tungsten' and 'products' predications from the context sentence.

The automatic methods we explored for the addition of rMRS structures corresponding to the antecedent and increasingly larger amounts of the context sentence underperformed manual integration of the textual antecedent. For +full-rMRS we expected a lower MRR because often the merging of irrelevant rMRSs from the context causes the correct target sentence to be ranked lower or not selected. The optimal amount of rMRS from an anaphorically linked sentence to add to that of the target sentence lies somewhere between the full rMRS and the elementary predication for the head of the antecedent. The main factor in the poor performance of partial rMRS integration is a consequence of the high degree of underspecification in the rMRSs output by our current QA system.

8 Conclusions

We have demonstrated experimentally that anaphora resolution is highly-relevant to open-domain QA (both in theory and in practice). However, the accuracy and manner of integration of domain-independent anaphora resolution is critical to effective deployment in open-domain QA. Using an extant resolver to conservatively enrich the rMRS of a target answer sentence containing a pronoun leads to an improvement in MRR, but there is still room for improvement as the approach fails to effectively enrich the representation in 29% of contexts. We have noted failures in the LL algorithm that could be addressed and in addition, the RASP system and other robust parsers (e.g. Buchholz (2002); Clark et al. (2002)) continue to be developed and improve, yielding more accurate starting points for anaphora resolution. Optimally defining the rMRS substructure from an anapohrically-linked context sentence to integrate with the target sentence rMRS is difficult in our current QA system. Extraction of more informative rMRS from the RASP system output should ameliorate this problem.

It is difficult to assess how potentially useful or genuinely effective domain-independent anaphora resolution would be for QA on different data.[4] Newspaper articles, as used in many extant TREC QA competitions, contain many anaphors. However, in different genres, such as scientific articles where we might expect QA systems to find genuinely useful application (e.g. Zweigenbaum (2003)), current anaphora resolution technology may not be so helpful. In this genre, sentences tend to be longer and a higher proportion of anaphors are definite descriptions, requiring more domain-dependent lexical information for high precision resolution.

Morton (2000) reports a small improvement when adding a coreference component to a QA system on the TREC 8 QA dataset. However, his results don't quantify the effect of coreference resolution effectively as his baseline system heuristically includes terms from surrounding sentences, being based on passage retrieval rather than sentence parsing and matching. Morton utilizes a supervised approach to pronominal anaphora resolution which, unlike ours, requires labelled training data and is arguably less domain-independent. He also resolves proper noun coreference and some def-

[4] It is easy to show such an effect in principle: Katz and Lin (2003) demonstrate that there are combinations of questions and document sets containing candidate answers that require syntactic analysis to recover GRs for accurate QA. With 16 carefully chosen questions which exhibit high semantic confusability between subjects and objects of the relevant predicates, system precision leapt from 29% for keyword matching to 84% for GR extraction.

inite description anaphora. Our future work will explore a more graded approach to non-coreference constraints by adopting a probabilistic framework of the type utilized by Ge et al. (1998), incorporating other domain-independent factors such as distance and mention in the resolution decision as well as extension of the approach to definite descriptions.

Acknowledgements

We thank Ann Copestake and Simone Teufel for their work setting up the task, evaluation and rMRS output from the parsers, and one anonymous referee for useful feedback which helped improve the final version.

References

E. J. Briscoe and J. Carroll. 2002. Robust accurate statistical annotation of general text. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation*, pages 1499–1504.

E. J. Briscoe, J. Carroll, J. Graham, and A. Copestake. 2002. Relational evaluation schemes. In *Proceedings of the beyond PARSEVAL Workshop at the 3rd International Conference on Language Resources and Evaluation*, pages 4–8.

S. Buchholz. 2002. *Memory-Based Grammatical Relation Finding*. Ph.D. thesis, University of Tilburg.

J. Carroll, G. Minnen, and E. J. Briscoe. 2003. Parser evaluation using a grammatical relation annotation scheme. In A. Abeille, editor, *Treebanks: Building and Using Syntactically Annotated Corpora*. Dordrecht:Kluwer.

J. Castano, J. Zhang, and J. Pustejovsky. 2002. Anaphora resolution in biomedial literature. In *Proceedings of the International Symposium on Reference Resolution*.

E. Charniak. 2000. A maximum-entropy-inspired parser. In *Proceedings of NAACL-2000*, pages 132–139.

S. Clark, J. Hockenmaier, and M. Steedman. 2002. Building deep dependency structures with a wide-coverage CCG parser. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 327–334.

M. Collins. 1997. Three generative, lexicalized models for statistical parsing. In *Proceedings of the 35th Annual Meeting of the ACL (jointly with the 8th Conference of the EACL)*, pages 16–23.

A. Copestake and D. Flickinger. 2000. An open-source grammar development environment and broad-coverage English grammar using HPSG. In *Proceedings of the Second conference on Language Resources and Evaluation*.

A. Copestake, D. Flickinger, C. J. Pollard, and I. A. Sag. 1999. Minimal recursion semantics: An introduction.

N. Ge, J. Hale, and E. Charniak. 1998. A statistical approach to anaphora resolution. In *Proceedings of the Sixth Workshop on Very Large Corpora*, pages 161–170.

B. Katz and J. Lin. 2003. Selectively using relations to improve precision in question answering. In *Proceedings of the EACL03 Workshop on NLP for QA*, pages 43–50.

C. Kennedy and B. Boguraev. 1996. Anaphora for everyone: Pronominal anaphora resolution without a parser. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING'96)*, pages 113–118.

S. Lappin and H. Leass. 1994. An algorithm for pronominal anaphora resolution. *Computational Linguistics*, 20(4):535–561.

S. Lappin and M. McCord. 1990. A syntactic filter on pronominal anaphora for slot grammar. In *Proceedings of 28th ACL*, pages 135–142.

G. Leech. 1992. 100 million words of English: the British National Corpus. *Language Research*, 28(1):1–13.

T. Morton. 2000. Coreference for NLP applications. In *Proceedings of 38th ACL*.

J. Preiss and E. Briscoe. 2003. Shallow or full parsing for anaphora resolution? An experiment with the Lappin and Leass algorithm. In *Proceedings of the Workshop on Anaphora Resolution*, pages 1–6.

J. Preiss. 2002. Choosing a parser for anaphora resolution. In *Proceedings of DAARC*, pages 175–180.

P. Zweigenbaum. 2003. Question answering in biomedicine. In *Proceedings of the EACL03 Workshop on NLP for QA*, pages 1–5.