# Using Mention Accessibility to Improve Coreference Resolution

**Kellie Webster and Joel Nothman**
School of Information Technologies
University of Sydney
NSW 2006, Australia
{kellie.webster, jnothman}@sydney.edu.au

## Abstract

Modern coreference resolution systems require linguistic and general knowledge typically sourced from costly, manually curated resources. Despite their intuitive appeal, results have been mixed. In this work, we instead implement fine-grained surface-level features motivated by cognitive theory. Our novel fine-grained feature specialisation approach significantly improves the performance of a strong baseline, achieving state-of-the-art results of 65.29 and 61.13% on CoNLL-2012 using gold and automatic preprocessing, with system extracted mentions.

## 1 Introduction

Coreference resolution (Pradhan et al., 2011, 2012) is the task of clustering mentions in a document according to their referent. For instance, we need to resolve Ehud Barak, his, and he as coreferential to understand the meaning of the excerpt:

**Israeli Prime Minister Ehud Barak** called **his** cabinet into special session late Wednesday , to discuss what **he** called a grave escalation of the level of violence ...

While knowledge-poor approaches establish a reasonable baseline, they perform poorly when positional and surface form heuristics break down. To address this, research has extracted world knowledge from manually curated resources including Wikipedia, Yago, Freebase, and FrameNet (e.g. Uryupina et al., 2011; Rahman and Ng, 2011; Ratinov and Roth, 2012; Hajishirzi et al., 2013; Durrett and Klein, 2014). Despite their intuitive appeal, results have been mixed. We instead focus on linguistic knowledge which can be extracted completely automatically, guided by insights from Accessibility theory (Ariel, 2001). This result is consistent with Wiseman et al. (2015) which similarly finds performance gains above state-of-the-art from extending simple, surface-level features.

We implement a mention classification scheme based on the Accessibility hierarchy and use this for feature specialisation, yielding state-of-the-art results of 65.29 and 61.13% on CoNLL-2012 on gold and automatic preprocessing, with system extracted mentions. Our approach is simple and effective, contributing to arguments for incorporating cognitive insights in computational modelling.

## 2 Accessibility Hierarchy

Accessibility theory (Ariel, 2001) builds on a body of cognitively motivated theories of discourse processing, notably Centering Theory (Grosz et al., 1995). Where Centering describes pronoun interpretation in terms of relative discourse entity salience, Accessibility theory expands from this, describing discourse entities as having corresponding human memory nodes which fluctuate in their degree of activation as the entity features in a discourse. The surface form of a reference indicates to the hearer how activated its corresponding node is expected to be. That is, surface form is an instruction for how to retrieve suitable referents, guiding the resolution of coreference. Relative degree of activation is captured in the theory's hierarchy of reference expression types, reproduced in Figure 1. Section 4 proposes a mapping of this hierarchy (derived for spoken Hebrew) to written English.

The hierarchy encodes and expands the widely-used rule of thumb that full names introduce an entity (their referent has low accessibility; it has not yet been discussed) and pronouns are anaphoric (their referent is a highly accessible, active dis-



*Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 432–437,
Berlin, Germany, August 7-12, 2016. ©2016 Association for Computational Linguistics

*Full name + modifier* < *Full name* < *Long definite description* < *Short definite description* < *Last name* < *First name* < *Distal demonstrative + modifier* < *Proximate demonstrative + modifier* < *Distal demonstrative + NP* < *Proximate demonstrative + NP* < *Distal demonstrative* < *Proximate demonstrative* < *Stressed pronoun + gesture* < *Stressed pronoun* < *Unstressed pronoun* < *Cliticised pronoun* < *Verbal inflections* < *Zero*

Figure 1: Accessibility hierarchy from Ariel (2001)

course entity); the accessibility of definite descriptions is intermediate. In this work, we show that the fine-grained categorisation in the Accessibility hierarchy can be leveraged to improve the discriminative power of a strong system, compared to coarser-grained typologies from previous work. That is, this work contributes valuable empirical support for the psycholinguistic theory.

## 3 Related Work

A particularly successful way to leverage mention classification has been to specialise modelling by mention type. Denis and Baldridge (2008) learn five different models, one each for proper name, definite nominal, indefinite nominal, third person pronoun, and non-third person pronoun. Bengtson and Roth (2008) and Durrett and Klein (2013) implement specialisation at the level of features within a model, rather than explicitly learning separate models. Bengtson and Roth (2008) prefix each base feature generated with the type of the current mention, one of proper name, nominal, or pronoun, for instance `nominal-head_match:true`. Durrett and Klein (2013) extend from this by learning a model over three versions of each base feature: unprefixed, conjoined with the type of the current mention, and conjoined with concatenation of the types of the current mention and candidate antecedent mention: `nominal+nominal-head_match=true`.

The success of Durrett and Klein is possible due to the large training dataset provided by OntoNotes (Pradhan et al., 2007). In this work, we successfully extend data-driven specialisation still further: Section 4 shows how we can discover fine-grained patterns in reference expression usage, and Section 5 how these patterns can be used to significantly improve the performance of a strong coreference system.

## 4 Accessibility Transitions in OntoNotes

In this section, we propose an implementation of the Accessibility hierarchy for written English and how this can be used to encode fine-grained discourse transitions. We discover trends in OntoNotes, over mentions automatically extracted from the DEV portion of English CoNLL-2012 (Pradhan et al., 2011).

| $AR$ | Description | % |
|---|---|---|
| 1 | Multi-word name + modifier | 7.7 |
| 2 | Multi-word name | 8.7 |
| 3 | Long indefinite description | 18.9 |
| 4 | Short indefinite description | 16.3 |
| 5 | Long definite description | 10.2 |
| 6 | Short definite description | 5.0 |
| 7 | Single-word name | 8.8 |
| 8 | Distal demonstrative + modifier | 0.2 |
| 9 | Proximate demonstrative + modifier | 0.0 |
| 10 | Distal demonstrative + NP | 0.7 |
| 11 | Proximate demonstrative + NP | 1.2 |
| 12 | Distal demonstrative | 0.8 |
| 13 | Proximate demonstrative | 0.5 |
| 14 | Pronoun | 21.0 |
| - | Zero | - |

Table 1: Accessibility rank values used in our experiments, with their base distribution over extracted NPs.

### 4.1 Mention classification

Our experiments start by classifying a mention's Accessibility rank value, $AR$. Table 1 gives the schema we propose for written English, along with the base distribution over extracted mentions. This mapping is a simple ordinal numbering of Figure 1 with the following refinements.

We have generalised last name and first name to single-word name ($AR$ = 7) and full name to multi-word name ($AR$ = 2) to handle non-person entities. Name modifiers are tokens without the head NER label, excluding determiners, possessive markers, and punctuation. We have introduced indefinite descriptions above definite descriptions since they are more likely to introduce discourse entities than definite descriptions are. We label any nominal started by the or a possessive pronoun as a definite; otherwise it is indefinite. Long descriptions comprise more than one token when possessive markers, punctuation, and articles are excluded. Distals start with those or that while

| *AR*(*antecedent*) \ *AR*(*anaphor*) | | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 10 | 11 | 12 | 13 | 14 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Name + modifier | 1 | 0.12 | 0.22 | 0.06 | | | | | 0.15 | | | | | | 0.48 |
| Multi-word name | 2 | 0.12 | | 0.31 | | | | 0.06 | 0.14 | | | | | | 0.40 |
| Long indefinite description | 3 | 0.21 | | | 0.07 | | 0.09 | 0.14 | | | | | | | 0.52 |
| Short indefinite description | 4 | 0.06 | | | | 0.05 | | 0.12 | 0.05 | | | | | | 0.65 |
| Long definite description | 5 | 0.14 | | | | | 0.21 | 0.15 | 0.09 | | | | | | 0.41 |
| Short definite description | 6 | 0.08 | | | | | 0.07 | 0.37 | 0.07 | | | | | | 0.39 |
| Single-word name | 7 | 0.15 | | | | | | | 0.49 | | | | | | 0.42 |
| Distal demonstrative + modifier | 8 | 0.01 | | | 0.05 | | | 0.05 | | 0.05 | | | | | 0.79 |
| Distal demonstrative + NP | 10 | 0.01 | | | | | 0.07 | 0.10 | | | 0.13 | | | | 0.54 |
| Proximate demonstrative + NP | 11 | 0.02 | | | | | 0.05 | 0.10 | 0.11 | | | 0.12 | | | 0.54 |
| Distal demonstrative | 12 | 0.00 | | | | | | 0.05 | 0.05 | | | | 0.34 | | 0.43 |
| Proximate demonstrative | 13 | 0.00 | | | | | | | 0.08 | | | 0.05 | | 0.05 | 0.71 |
| Pronoun | 14 | 0.08 | | | | | | | 0.09 | | | | | | 0.82 |

Table 2: Accessibility transitions (>0.05) CoNLL-2012 DEV.

proximates start with these or this.

### 4.2 Discourse Transitions

Discourse transitions are then $AR$ tuples whose values come from mentions aligned to the same gold cluster. We chose 2-tuples, whose values come from mention-antecedent pairs, since mention-pair models have dominated the research space. However, we generate up to three pairs per mention since antecedents are latent at the entity level. That is, for he in the following, we generate pairs (1, 14) and (14, 14).

**Israeli Prime Minister Ehud Barak**$_{AR=1}$ called **his**$_{AR=14}$ cabinet into special session late Wednesday , to discuss what **he**$_{AR=14}$ called a grave escalation of the level of violence ...

The aggregated counts for each pair type are represented in Table 2, with $AR(antecedent)$ on the vertical and $AR(anaphor)$ on the horizontal. The first column gives the proportion of cluster-initial mentions of each $AR$ type (e.g. 21% of gold clusters have a long indefinite description as their first mention). Each row is normalised to sum to 1 so each row indicates the probability distribution for the expected next mention of a cluster. For clarity, only values 0.05 and higher are shown.

We can see that commonly used rules of thumb are borne out in this data, though with some extra granularity. Modified and multi-word names reduce to single-word names, and both reduce to pronouns. Single word names retain their mention form and reduce to pronouns with roughly equal probability. All mention types reduce to be pronouns and, once reference has reduced to be pronominal, there is a high likelihood (82%) that this form will be retained.

Encouragingly, we can also see transitions in Table 2 can not be expressed with the coarser-grained typologies of prior work. Firstly, mention article is important. Long indefinite descriptions are more likely to start coreference clusters than long definite descriptions (21% vs. 14%), which are in turn much more likely to start clusters than demonstratives. Mention length is also important: short indefinite descriptions are more likely to reduce to pronouns than long definite descriptions and short definite descriptions have a higher chance of being retained throughout the discourse than long definite descriptions. Exploring further, of coreferential pairs where both mentions are short definite descriptions, 86% are head matched, compared to 60% of long definite descriptions; 60% of short definite descriptions are string matched, compared to 27% of long.

| *AR* | Description | |
|---|---|---|
| 1 | Name + modifier | 0.56 |
| 2 | Multi-word name | 0.65 |
| 3 | Long indefinite description | 0.89 |
| 4 | Short indefinite description | 0.92 |
| 5 | Long definite description | 0.75 |
| 6 | Short definite description | 0.54 |
| 7 | Single-word name | 0.44 |
| 8 | Distal demonstrative + modifier | 0.69 |
| 9 | Proximate demonstrative + modifier | 1.00 |
| 10 | Distal demonstrative + NP | 0.43 |
| 11 | Proximate demonstrative + NP | 0.41 |
| 12 | Distal demonstrative | 0.43 |
| 13 | Proximate demonstrative | 0.60 |
| 14 | Pronoun | 0.21 |

Table 3: Proportion of singletons by $AR$.

### 4.3 Anaphoricity

Table 3 gives the proportion of extracted mentions which can not be aligned to gold mentions, by $AR$ value. Modelling these discourse singletons is important for models jointly learning coreference and anaphoricity (Webster and Curran, 2014).

| | Gold | | | | Auto | | | |
|---|---|---|---|---|---|---|---|---|
| | MUC | $B^3$ | CEAFE | CoNLL | MUC | $B^3$ | CEAFE | CoNLL |
| Fernandes et al. (2012) | 72.18 | 59.17 | 55.72 | 62.36 | 70.51 | 57.58 | 53.86 | 60.65 |
| Björkelund and Kuhn (2014) | 73.80 | 62.00 | 59.06 | 64.95 | 70.72 | 58.58 | 55.61 | 61.63 |
| LIMERIC Baseline | 74.07 | 60.91 | 58.57 | 64.52 | 70.36 | 56.60 | 54.42 | 60.46 |
| + Fine-Grained Specialisation | *74.73* | *61.72* | *59.43* | **65.29** | 70.72 | *57.40* | *55.26* | **61.13** |

Table 4: Performance on CoNLL-2012 TEST evaluated with gold and automatic annotations and system extracted mentions.

After pronouns, demonstratives and proper names have low proportion of singletons. Single word names are less likely to be singletons than modified and multi-word names. We highlight two contributing factors. The first is that certain names, particularly the children of an apposition, are not markable in OntoNotes. The second is that the burden of supplying disambiguation will be more worthwhile for important entities.

Consistent with Recasens et al. (2013), indefinites are more likely to be singletons than definites, and long definites are more likely than short definites. Since length and article are the key factors for $AR$ typing, this is good evidence in favour of using the hierarchy's fine-grained classification.

## 5 Experiments

In this section, we show how fine-grained feature specialisation can significantly improve the performance of LIMERIC, a competitive coreference resolution system. This strength demonstrates that simple surface-form features have yet to be fully utilised in current modelling, and that cognitive theory can guide their development.

### 5.1 LIMERIC

The system we base our work on is LIMERIC (Webster and Curran, 2014). We choose this system due to its cognitive motivation and strong performance. Importantly, this system already uses the coarse-grained featurisation of Durrett and Klein (2013), allowing us to directly measure the impact of our proposed fine-grained featurisation.

We, however, improve it in a number of ways. The biggest performance boosts came from using MIRA (Margin Infused Relaxation Algorithm) updates in place of standard perceptron updates and implementing the full range of common features from the literature. We also fix a number of bug fixes and improve mention extraction. This improved system forms our LIMERIC baseline in Table 4.

### 5.2 Fine-Grained Feature Specialisation

We build on work in discourse transition prefixing (particularly Durrett and Klein, 2013), which expands the feature space of a learner by including multiple versions of each generated feature. LIMERIC previously used three versions of each feature: one unprefixed, one prefixed with the current mention's type (one of name, nominal, or pronoun), and one prefixed with the concatenation of the types of the current and candidate antecedents. In this work, we introduce a fourth prefix, formed by concatenating the $AR$ of the current mention with that of the closest mention in the candidate antecedent cluster.

The power of such transition features is that they allow us to learn, for instance, that pronoun to name transitions are preferred when the anaphor is distant from its antecedent and the name mention is one token, or that head match is a particularly strong indicator of coreferentiality between short definite nominals: `6+6-head_match=true`.

### 5.3 Results

Table 4 tabulates system performance on CoNLL-2012 TEST using system extracted mentions and v8.01 of the official scorer (Pradhan et al., 2014).

Comparing feature specialisation against the LIMERIC baseline, we can see that it yields substantial performance gains on all metrics and both evaluation settings. Performance gains indicated in bold are statistically significant for the conservative p = 0.01 using bootstrap re-sampling[1]. Performance gains indicated in italics are significant at the standard threshold of p = 0.05.

We benchmark against the state-of-the-art by

[1]Since Specialisation is a development of LIMERIC, the two models are not independent which means we would expect to see relatively high confidence values for relatively small gains in score (see Berg-Kirkpatrick et al., 2012).

comparing performance to the winner of the shared task (Fernandes et al., 2012), as well as the best documented system at the time of this work (Björkelund and Kuhn, 2014). Fine-grained feature specialisation improves LIMERIC's performance to push past that of Björkelund and Kuhn (2014) when using gold preprocessing. Furthermore, on the difficult automatic setting, we outperform Fernandes et al. (2012) and are not significantly worse than Björkelund and Kuhn (2014).

On the link-based MUC and $B^3$ metrics, our recall gains are larger than our precision gains. That is, specialisation enables coreference indicators to accrue sufficient weight so as to promote new coreference links, a known problem case for modern systems. We found particularly enhanced weight on features for relaxed string matching.

## 6 Conclusion

In this paper, we have found fine-grained patterns in reference expression usage based on the Accessibility hierarchy and shown how these can be used to significantly improve the performance of a strong system, LIMERIC. Despite being simple to implement, we achieve comparable or improved performance than the best reported results, furthering arguments for incorporating cognitive insights in computational modelling.

## 7 Acknowledgements

The authors thank their anonymous reviewers and members of the Schwa Lab at the University of Sydney for their insightful and helpful feedback. The first author was supported by an Australian Postgraduate Award scholarship.

## References

Mira Ariel. 2001. Accessibility theory: An overview. *Text representation: Linguistic and psycholinguistic aspects*, pages 29–87.

Eric Bengtson and Dan Roth. 2008. Understanding the value of features for coreference resolution. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 294–303. Association for Computational Linguistics.

Taylor Berg-Kirkpatrick, David Burkett, and Dan Klein. 2012. An empirical investigation of statistical significance in nlp. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 995–1005. Association for Computational Linguistics, Jeju Island, Korea.

Anders Björkelund and Jonas Kuhn. 2014. Learning structured perceptrons for coreference resolution with latent antecedents and non-local features. *ACL, Baltimore, MD, USA, June*.

Pascal Denis and Jason Baldridge. 2008. Specialized models and ranking for coreference resolution. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 660–669. Association for Computational Linguistics.

Greg Durrett and Dan Klein. 2013. Easy victories and uphill battles in coreference resolution. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Greg Durrett and Dan Klein. 2014. A joint model for entity analysis: Coreference, typing, and linking. In *Proceedings of the Transactions of the Association for Computational Linguistics*.

Eraldo Fernandes, Cícero dos Santos, and Ruy Milidiú. 2012. Latent structure perceptron with feature induction for unrestricted coreference resolution. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 41–48. Association for Computational Linguistics, Jeju Island, Korea.

Barbara J Grosz, Scott Weinstein, and Aravind K Joshi. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225.

Hannaneh Hajishirzi, Leila Zilles, Daniel S Weld, and Luke Zettlemoyer. 2013. Joint coreference resolution and named-entity linking with multi-pass sieves. pages 289–299.

Sameer Pradhan, Xiaoqiang Luo, Marta Recasens, Eduard Hovy, Vincent Ng, and Michael Strube. 2014. Scoring coreference partitions of predicted mentions: A reference implementation.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. Conll-2012 shared task: Modeling multilingual unrestricted coreference in ontonotes. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 1–40. Association for Computational Linguistics, Jeju Island, Korea.

Sameer Pradhan, Lance Ramshaw, Mitchell Marcus, Martha Palmer, Ralph Weischedel, and Nianwen Xue. 2011. Conll-2011 shared task: Modeling unrestricted coreference in ontonotes. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–27. Association for Computational Linguistics, Portland, Oregon, USA.

Sameer S. Pradhan, Eduard H. Hovy, Mitchell P. Marcus, Martha Palmer, Lance A. Ramshaw, and Ralph M. Weischedel. 2007. Ontonotes: a unified relational semantic representation. *Int. J. Semantic Computing*, 1(4):405–419.

Altaf Rahman and Vincent Ng. 2011. Coreference resolution with world knowledge. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pages 814–824.

Lev Ratinov and Dan Roth. 2012. Learning-based multi-sieve co-reference resolution with knowledge. In *EMNLP*.

Marta Recasens, Marie-Catherine de Marneffe, and Christopher Potts. 2013. The life and death of discourse entities: Identifying singleton mentions. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 627–633. Association for Computational Linguistics, Atlanta, Georgia.

Olga Uryupina, Massimo Poesio, Claudio Giuliano, and Kateryna Tymoshenko. 2011. Disambiguation and filtering methods in using web knowledge for coreference resolution. In *Proceedings of the 24th International Florida Artificial Intelligence Research Society Conference*, pages 317–322.

Kellie Webster and James R Curran. 2014. Limited memory incremental coreference resolution. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2129–2139. Dublin,Ireland.

Sam Wiseman, Alexander M Rush, Stuart M Shieber, Jason Weston, Heather Pon-Barry, Stuart M Shieber, Nicholas Longenbaugh, Sam Wiseman, Stuart M Shieber, Elif Yamangil, et al. 2015. Learning anaphoricity and antecedent ranking features for coreference resolution. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 92–100. Association for Computational Linguistics.