

Coreference-inspired Coherence Modeling

Micha Elsner and Eugene Charniak

Brown Laboratory for Linguistic Information Processing (BLLIP)

Brown University

Providence, RI 02912

{melsner, ec}@cs.brown.edu

Abstract

Research on coreference resolution and summarization has modeled the way entities are realized as concrete phrases in discourse. In particular there exist models of the noun phrase syntax used for discourse-new versus discourse-old referents, and models describing the likely distance between a pronoun and its antecedent. However, models of discourse coherence, as applied to information ordering tasks, have ignored these kinds of information. We apply a discourse-new classifier and pronoun coreference algorithm to the information ordering task, and show significant improvements in performance over the entity grid, a popular model of local coherence.

1 Introduction

Models of discourse coherence describe the relationships between nearby sentences, in which previous sentences help make their successors easier to understand. Models of coherence have been used to impose an order on sentences for multidocument summarization (Barzilay et al., 2002), to evaluate the quality of human-authored essays (Miltsakaki and Kukich, 2004), and to insert new information into existing documents (Chen et al., 2007).

These models typically view a sentence either as a bag of words (Foltz et al., 1998) or as a bag of entities associated with various syntactic roles (Lapata and Barzilay, 2005). However, a mention of an entity contains more information than just its head and syntactic role. The referring expression itself contains discourse-motivated information distinguishing familiar entities from unfamiliar and salient from

non-salient. These patterns have been studied extensively, by linguists (Prince, 1981; Fraurud, 1990) and in the field of coreference resolution. We draw on the coreference work, taking two standard models from the literature and applying them to coherence modeling.

Our first model distinguishes discourse-new from discourse-old noun phrases, using features based on Uryupina (2003). Discourse-new NPs are those whose referents have not been previously mentioned in the discourse. As noted by studies since Hawkins (1978), there are marked syntactic differences between the two classes.

Our second model describes pronoun coreference. To be intelligible, pronouns must be placed close to appropriate referents with the correct number and gender. Centering theory (Grosz et al., 1995) describes additional constraints about which entities in a discourse can be pronominalized: if there are pronouns in a segment, they must include the backward-looking center. We use a model which probabilistically attempts to describe these preferences (Ge et al., 1998).

These two models can be combined with the entity grid described by Lapata and Barzilay (2005) for significant improvement. The magnitude of the improvement is particularly interesting given that Barzilay and Lapata (2005) do use a coreference system but are unable to derive much advantage from it.

2 Discourse-new Model

In the task of discourse-new classification, the model is given a referring expression (as in previous work, we consider only NPs) from a document and must

determine whether it is a first mention (*discourse-new*) or a subsequent mention (*discourse-old*). Features such as full names, appositives, and restrictive relative clauses are associated with the introduction of unfamiliar entities into discourse (Hawkins, 1978; Fraurud, 1990; Vieira and Poesio, 2000). Classifiers in the literature include (Poesio et al., 2005; Uryupina, 2003; Ng and Cardie, 2002). The system of Nenkova and McKeown (2003) works in the opposite direction. It is designed to rewrite the references in multi-document summaries, so that they conform to the common discourse patterns.

We construct a maximum-entropy classifier using syntactic and lexical features derived from Uryupina (2003), and a publicly available learning tool (Daumé III, 2004). Our system scores 87.4% (F-score of the *disc-new* class on the MUC-7 formal test set); this is comparable to the state-of-the-art system of Uryupina (2003), which scores 86.9¹.

To model coreference with this system, we assign each NP in a document a label $L_{np} \in \{new, old\}$. Since the correct labeling depends on the coreference relationships between the NPs, we need some way to guess at this; we take all NPs with the same head to be coreferent, as in the non-coreference version of (Barzilay and Lapata, 2005)². We then take the probability of a document as $\prod_{np: NPs} P(L_{np}|np)$.

We must make several small changes to the model to adapt it to this setting. For the discourse-new classification task, the model’s most important feature is whether the head word of the NP to be classified has occurred previously (as in Ng and Cardie (2002) and Vieira and Poesio (2000)). For coherence modeling, we must remove this feature, since it depends on document order, which is precisely what we are trying to predict. The coreference heuristic will also fail to resolve any pronouns, so we discard them.

Another issue is that NPs whose referents are familiar tend to resemble discourse-old NPs, even though they have not been previously mentioned (Fraurud, 1990). These include unique objects like *the FBI* or generic ones like *danger* or *percent*. To

¹Poesio et al. (2005) score 90.2%, but on a different corpus.

²Unfortunately, this represents a substantial sacrifice; as Poesio and Vieira (1998) show, only about 2/3 of definite descriptions which are anaphoric have the same head as their antecedent.

avoid using these deceptive phrases as examples of discourse-newness, we attempt to heuristically remove them from the training set by discarding any NP whose head occurs only once in the document³.

The labels we apply to NPs in our test data are systematically biased by the “same head” heuristic we use for coreference. This is a disadvantage for our system, but it has a corresponding advantage—we can use training data labeled using the same heuristic, without any loss in performance on the coherence task. NPs we fail to learn about during training are likely to be mislabeled at test time anyway, so performance does not degrade by much. To counter this slight degradation, we can use a much larger training corpus, since we no longer require gold-standard coreference annotations.

3 Pronoun Coreference Model

Pronoun coreference is another important aspect of coherence— if a pronoun is used too far away from any natural referent, it becomes hard to interpret, creating confusion. Too many referents, however, create ambiguity. To describe this type of restriction, we must model the probability of the text containing pronouns (denoted r_i), jointly with their referents a_i . (This takes more work than simply resolving the pronouns conditioned on the text.) The model of Ge et al. (1998) provides the requisite probabilities:

$$P(a_i, r_i | a_i^{i-1}) = P(a_i | h(a_i), m(a_i)) \\ P_{gen}(a_i, r_i) P_{num}(a_i, r_i)$$

Here $h(a)$ is the Hobbs distance (Hobbs, 1976), which measures distance between a pronoun and prospective antecedent, taking into account various factors, such as syntactic constraints on pronouns. $m(a)$ is the number of times the antecedent has been mentioned previously in the document (again using “same head” coreference for full NPs, but also counting the previous antecedents a_i^{i-1}). P_{gen} and P_{num} are distributions over gender and number given words. The model is trained using a small hand-annotated corpus first used in Ge et al. (1998).

³Bean and Riloff (1999) and Uryupina (2003) construct quite accurate classifiers to detect unique NPs. However, some preliminary experiments convinced us that our heuristic method worked well enough for the purpose.

	Disc. Acc	Disc. F	Ins.
Random	50.00	50.00	12.58
Entity Grid	76.17	77.55	19.57
Disc-New	70.35	73.47	16.27
Pronoun	55.77	62.27	13.95
EGrid+Disc-New	78.88	80.31	21.93
Combined	79.60	81.02	22.98

Table 1: Results on 1004 WSJ documents.

Finding the probability of a document using this model requires us to sum out the antecedents a . Unfortunately, because each a_i is conditioned on the previous ones, this cannot be done efficiently. Instead, we use a greedy search, assigning each pronoun left to right. Finally we report the probability of the resulting sequence of pronoun assignments.

4 Baseline Model

As a baseline, we adopt the entity grid (Lapata and Barzilay, 2005). This model outperforms a variety of word overlap and semantic similarity models, and is used as a component in the state-of-the-art system of Soricut and Marcu (2006). The entity grid represents each entity by tracking the syntactic roles in which it appears throughout the document. The internal syntax of the various referring expressions is ignored. Since it also uses the “same head” coreference heuristic, it also disregards pronouns.

Since the three models use very different feature sets, we combine them by assuming independence and multiplying the probabilities.

5 Experiments

We evaluate our models using two tasks, both based on the assumption that a human-authored document is coherent, and uses the best possible ordering of its sentences (see Lapata (2006)). In the discrimination task (Barzilay and Lapata, 2005), a document is compared with a random permutation of its sentences, and we score the system correct if it indicates the original as more coherent⁴.

⁴Since the model might refuse to make a decision by scoring a permutation the same as the original, we also report F-score, where precision is $correct/decisions$ and recall is $correct/total$.

Discrimination becomes easier for longer documents, since a random permutation is likely to be much less similar to the original. Therefore we also test our systems on the task of insertion (Chen et al., 2007), in which we remove a sentence from a document, then find the point of insertion which yields the highest coherence score. The reported score is the average fraction of sentences per document reinserted in their original position (averaged over documents, not sentences, so that longer documents do not disproportionately influence the results)⁵.

We test on sections 14-24 of the Penn Treebank (1004 documents total). Previous work has focused on the AIRPLANE corpus (Barzilay and Lee, 2004), which contains short announcements of airplane crashes written by and for domain experts. These texts use a very constrained style, with few discourse-new markers or pronouns, and so our system is ineffective; the WSJ corpus is much more typical of normal informative writing. Also unlike previous work, we do not test the task of completely reconstructing a document’s order, since this is computationally intractable and results on WSJ documents⁶ would likely be dominated by search errors.

Our results are shown in table 5. When run alone, the entity grid outperforms either of our models. However, all three models are significantly better than random. Combining all three models raises discrimination performance by 3.5% over the baseline and insertion by 3.4%. Even the weakest component, pronouns, contributes to the joint model; when it is left out, the resulting *EGrid + Disc-New* model is significantly worse than the full combination. We test significance using Wilcoxon’s signed-rank test; all results are significant with $p < .001$.

6 Conclusions

The use of these coreference-inspired models leads to significant improvements in the baseline. Of the two, the discourse-new detector is by far more effective. The pronoun model’s main problem is that, although a pronoun may have been displaced from its original position, it can often find another seemingly acceptable referent nearby. Despite this issue

⁵Although we designed a metric that distinguishes near misses from random performance, it is very well correlated with exact precision, so, for simplicity’s sake, we omit it.

⁶Average 22 sentences, as opposed to 11.5 for AIRPLANE.

it performs significantly better than chance and is capable of slightly improving the combined model. Both of these models are very different from the lexical and entity-based models currently used for this task (Soricut and Marcu, 2006), and are probably capable of improving the state of the art.

As mentioned, Barzilay and Lapata (2005) uses a coreference system to attempt to improve the entity grid, but with mixed results. Their method of combination is quite different from ours; they use the system's judgements to define the "entities" whose repetitions the system measures⁷. In contrast, we do not attempt to use any proposed coreference links; as Barzilay and Lapata (2005) point out, these links are often erroneous because the disordered input text is so dissimilar to the training data. Instead we exploit our models' ability to measure the probability of various aspects of the text.

Acknowledgements

Chen and Barzilay, reviewers, DARPA, et al.

References

- Regina Barzilay and Mirella Lapata. 2005. Modeling local coherence: an entity-based approach. In *ACL 2005*.
- Regina Barzilay and Lillian Lee. 2004. Catching the drift: Probabilistic content models, with applications to generation and summarization. In *HLT-NAACL 2004*, pages 113–120.
- Regina Barzilay, Noemie Elhadad, and Kathleen McKeown. 2002. Inferring strategies for sentence ordering in multidocument news summarization. *Journal of Artificial Intelligence Results (JAIR)*, 17:35–55.
- David L. Bean and Ellen Riloff. 1999. Corpus-based identification of non-anaphoric noun phrases. In *ACL'99*, pages 373–380.
- Erdong Chen, Benjamin Snyder, and Regina Barzilay. 2007. Incremental text structuring with online hierarchical ranking. In *Proceedings of EMNLP*.
- Hal Daumé III. 2004. Notes on CG and LM-BFGS optimization of logistic regression. Paper available at <http://pub.hal3.name#daume04cg-bfgs>, implementation available at <http://hal3.name/megam/>, August.
- Peter Foltz, Walter Kintsch, and Thomas Landauer. 1998. The measurement of textual coherence with latent semantic analysis. *Discourse Processes*, 25(2&3):285–307.
- Kari Fraurud. 1990. Definiteness and the processing of noun phrases in natural discourse. *Journal of Semantics*, 7(4):395–433.
- Niyu Ge, John Hale, and Eugene Charniak. 1998. A statistical approach to anaphora resolution. In *Proceedings of the Sixth Workshop on Very Large Corpora*, pages 161–171.
- Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225.
- John A. Hawkins. 1978. *Definiteness and indefiniteness: a study in reference and grammaticality prediction*. Croom Helm Ltd.
- Jerry R. Hobbs. 1976. Pronoun resolution. Technical Report 76-1, City College New York.
- Mirella Lapata and Regina Barzilay. 2005. Automatic evaluation of text coherence: Models and representations. In *IJCAI*, pages 1085–1090.
- Mirella Lapata. 2006. Automatic evaluation of information ordering: Kendall's tau. *Computational Linguistics*, 32(4):1–14.
- E. Miltsakaki and K. Kukich. 2004. Evaluation of text coherence for electronic essay scoring systems. *Nat. Lang. Eng.*, 10(1):25–55.
- Ani Nenkova and Kathleen McKeown. 2003. References to named entities: a corpus study. In *NAACL '03*, pages 70–72.
- Vincent Ng and Claire Cardie. 2002. Identifying anaphoric and non-anaphoric noun phrases to improve coreference resolution. In *COLING*.
- Massimo Poesio and Renata Vieira. 1998. A corpus-based investigation of definite description use. *Computational Linguistics*, 24(2):183–216.
- Massimo Poesio, Mijail Alexandrov-Kabadjov, Renata Vieira, Rodrigo Goulart, and Olga Uryupina. 2005. Does discourse-new detection help definite description resolution? In *Proceedings of the Sixth International Workshop on Computational Semantics*, Tillburg.
- Ellen Prince. 1981. Toward a taxonomy of given-new information. In Peter Cole, editor, *Radical Pragmatics*, pages 223–255. Academic Press, New York.
- Radu Soricut and Daniel Marcu. 2006. Discourse generation using utility-trained coherence models. In *ACL-2006*.
- Olga Uryupina. 2003. High-precision identification of discourse new and unique noun phrases. In *Proceedings of the ACL Student Workshop*, Sapporo.
- Renata Vieira and Massimo Poesio. 2000. An empirically-based system for processing definite descriptions. *Computational Linguistics*, 26(4):539–593.

⁷We attempted this method for pronouns using our model, but found it ineffective.