Extending the Entity Grid with Entity-Specific Features

Micha Elsner
School of Informatics
University of Edinburgh
melsner0@gmail.com

Eugene Charniak
Department of Computer Science
Brown University, Providence, RI 02912
ec@cs.brown.edu

Abstract

We extend the popular entity grid representation for local coherence modeling. The grid abstracts away information about the entities it models; we add discourse prominence, named entity type and coreference features to distinguish between important and unimportant entities. We improve the best result for WSJ document discrimination by 6%.

1 Introduction

A well-written document is coherent (Halliday and Hasan, 1976)– it structures information so that each new piece of information is interpretable given the preceding context. Models that distinguish coherent from incoherent documents are widely used in generation, summarization and text evaluation.

Among the most popular models of coherence is the entity grid (Barzilay and Lapata, 2008), a statistical model based on Centering Theory (Grosz et al., 1995). The grid models the way texts focus on important entities, assigning them repeatedly to prominent syntactic roles. While the grid has been successful in a variety of applications, it is still a surprisingly unsophisticated model, and there have been few direct improvements to its simple feature set. We present an extension to the entity grid which distinguishes between different types of entity, resulting in significant gains in performance[1].

At its core, the grid model works by predicting whether an entity will appear in the next sentence (and what syntactic role it will have) given its history of occurrences in the previous sentences. For instance, it estimates the probability that "Clinton" will be the subject of sentence 2, given that it was the subject of sentence 1. The standard grid model uses no information about the entity itself– the probability is the same whether the entity under discussion is "Hillary Clinton" or "wheat". Plainly, this assumption is too strong. Distinguishing important from unimportant entity types is important in coreference (Haghighi and Klein, 2010) and summarization (Nenkova et al., 2005); our model applies the same insight to the entity grid, by adding information from syntax, a named-entity tagger and statistics from an external coreference corpus.

2 Related work

Since its initial appearance (Lapata and Barzilay, 2005; Barzilay and Lapata, 2005), the entity grid has been used to perform wide variety of tasks. In addition to its first proposed application, sentence ordering for multidocument summarization, it has proven useful for story generation (McIntyre and Lapata, 2010), readability prediction (Pitler et al., 2010; Barzilay and Lapata, 2008) and essay scoring (Burstein et al., 2010). It also remains a critical component in state-of-the-art sentence ordering models (Soricut and Marcu, 2006; Elsner and Charniak, 2008), which typically combine it with other independently-trained models.

There have been few attempts to improve the entity grid directly by altering its feature representation. Filippova and Strube (2007) incorporate semantic relatedness, but find no significant improve-

[1] A public implementation is available via https://bitbucket.org/melsner/browncoherence.

Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics:shortpapers, pages 125–129,
Portland, Oregon, June 19-24, 2011. ©2011 Association for Computational Linguistics

1 [Visual meteorological conditions]$_S$ prevailed for [the personal cross country flight for which [a VFR flight plan]$_O$ was filed]$_X$.
2 [The flight]$_S$ originated at [Nuevo Laredo , Mexico]$_X$, at [approximately 1300]$_X$.

s	conditions	plan	flight	laredo
1	S	O	X	-
2	-	-	S	X

Figure 1: A short text (using NP-only mention detection), and its corresponding entity grid. The numeric token "1300" is removed in preprocessing.

ment over the original model. Cheung and Penn (2010) adapt the grid to German, where focused constituents are indicated by sentence position rather than syntactic role. The best entity grid for English text, however, is still the original.

3 Entity grids

The entity grid represents a document as a matrix (Figure 1) with a row for each sentence and a column for each entity. The entry for (sentence i, entity j), which we write $r_{i,j}$, represents the syntactic role that entity takes on in that sentence: subject (**S**), object (**O**), or some other role (**X**)[2]. In addition, there is a special marker (**-**) for entities which do not appear at all in a given sentence.

To construct a grid, we must first decide which textual units are to be considered "entities", and how the different mentions of an entity are to be linked. We follow the -COREFERENCE setting from Barzilay and Lapata (2005) and perform heuristic coreference resolution by linking mentions which share a head noun. Although some versions of the grid use an automatic coreference resolver, this often fails to improve results; in Barzilay and Lapata (2005), coreference improves results in only one of their target domains, and actually hurts for readability prediction. Their results, moreover, rely on running coreference on the document *in its original order*; in a summarization task, the correct order is not known, which will cause even more resolver errors.

To build a model based on the grid, we treat the columns (entities) as independent, and look at local transitions between sentences. We model the transitions using the generative approach given in Lapata and Barzilay (2005)[3], in which the model estimates the probability of an entity's role in the next sentence, $r_{i,j}$, given its history in the previous two sentences, $r_{i-1,j}, r_{i-2,j}$. It also uses a single entity-specific feature, salience, determined by counting the total number of times the entity is mentioned in the document. We denote this feature vector $F_{i,j}$. For example, the vector for "flight" after the last sentence of the example would be $F_{3,flight} = \langle X, S, sal = 2\rangle$. Using two sentences of context and capping salience at 4, there are only 64 possible vectors, so we can learn an independent multinomial distribution for each F. However, the number of vectors grows exponentially as we add features.

4 Experimental design

We test our model on two experimental tasks, both testing its ability to distinguish between correct and incorrect orderings for WSJ articles. In *document discrimination* (Barzilay and Lapata, 2005), we compare a document to a random permutation of its sentences, scoring the system correct if it prefers the original ordering[4].

We also evaluate on the more difficult task of *sentence insertion* (Chen et al., 2007; Elsner and Charniak, 2008). In this task, we remove each sentence from the article and test whether the model prefers to re-insert it at its original location. We report the average proportion of correct insertions per document.

As in Elsner and Charniak (2008), we test on sections 14-24 of the Penn Treebank, for 1004 test documents. We test significance using the Wilcoxon Sign-rank test, which detects significant differences in the medians of two distributions[5].

5 Mention detection

Our main contribution is to extend the entity grid by adding a large number of entity-specific features. Before doing so, however, we add non-head nouns to the grid. Doing so gives our feature-based model

[2] Roles are determined heuristically using trees produced by the parser of (Charniak and Johnson, 2005).

[3] Barzilay and Lapata (2005) give a discriminative model, which relies on the same feature set as discussed here.

[4] As in previous work, we use 20 random permutations of each document. Since the original and permutation might tie, we report both accuracy and balanced F-score.

[5] Our reported scores are means, but to test significance of differences in means, we would need to use a parametric test.

	Disc. Acc	Disc. F	Ins.
Random	50.0	50.0	12.6
Grid: NPs	74.4	76.2	21.3
Grid: all nouns†	**77.8**	**79.7**	**23.5**

Table 1: Discrimination scores for entity grids with different mention detectors on WSJ development documents. † indicates performance on both tasks is significantly different from the previous row of the table with p=.05.

more information to work with, but is beneficial even to the standard entity grid.

We alter our mention detector to add all nouns in the document to the grid[6], even those which do not head NPs. This enables the model to pick up premodifiers in phrases like "a **Bush** spokesman", which do not head NPs in the Penn Treebank. Finding these is also necessary to maximize coreference recall (Elsner and Charniak, 2010). We give non-head mentions the role **X**. The results of this change are shown in Table 1; discrimination performance increases about 4%, from 76% to 80%.

6 Entity-specific features

As we mentioned earlier, the standard grid model does not distinguish between different types of entity. Given the same history and salience, the same probabilities are assigned to occurrences of "Hillary Clinton", "the airlines", or "May 25th", even though we know *a priori* that a document is more likely to be about Hillary Clinton than it is to be about May 25th. This problem is exacerbated by our same-head coreference heuristic, which sometimes creates spurious entities by lumping together mentions headed by nouns like "miles" or "dollars". In this section, we add features that separate important entities from less important or spurious ones.

Proper Does the entity have a proper mention?

Named entity The majority OPENNLP Morton et al. (2005) named entity label for the coreferential chain.

Modifiers The total number of modifiers in all mentions in the chain, bucketed by 5s.

Singular Does the entity have a singular mention?

News articles are likely to be about people and organizations, so we expect these named entity tags, and proper NPs in general, to be more important to the discourse. Entities with many modifiers throughout the document are also likely to be important, since this implies that the writer wishes to point out more information about them. Finally, singular nouns are less likely to be generic.

We also add some features to pick out entities that are likely to be spurious or unimportant. These features depend on in-domain coreference data, but they do not require us to run a coreference resolver on the target document itself. This avoids the problem that coreference resolvers do not work well for disordered or automatically produced text such as multidocument summary sentences, and also avoids the computational cost associated with coreference resolution.

Linkable Was the head word of the entity ever marked as coreferring in MUC6?

Unlinkable Did the head word of the entity occur 5 times in MUC6 and never corefer?

Has pronouns Were there 5 or more pronouns coreferent with the head word of the entity in the NANC corpus? (Pronouns in NANC are automatically resolved using an unsupervised model (Charniak and Elsner, 2009).)

No pronouns Did the head word of the entity occur over 50 times in NANC, and have fewer than 5 coreferent pronouns?

To learn probabilities based on these features, we model the conditional probability $p(r_{i,j}|F)$ using multilabel logistic regression. Our model has a parameter for each combination of syntactic role r, entity-specific feature h and feature vector F: $r \times h \times F$. This allows the old and new features to interact while keeping the parameter space tractable[7].

In Table 2, we examine the changes in our estimated probability in one particular context: an entity with salience 3 which appeared in a non-emphatic role in the previous sentence. The standard entity grid estimates that such an entity will be the subject of the next sentence with a probability of about

[6] Barzilay and Lapata (2008) uses NPs as mentions; we are unsure whether all other implementations do the same, but we believe we are the first to make the distinction explicit.

[7] We train the regressor using OWLQN (Andrew and Gao, 2007), modified and distributed by Mark Johnson as part of the Charniak-Johnson parse reranker (Charniak and Johnson, 2005).

Context	P(next role is subj)
Standard egrid	.045
Head coref in MUC6	.013
...and proper noun	.025
...and NE type person	.037
...and 5 modifiers overall	.133
Never coref in MUC6	.006
...and NE type date	.001

Table 2: Probability of an entity appearing as subject of the next sentence, given the history **- X, salience 3**, and various entity-specific features.

.04. For most classes of entity, we can see that this is an overestimate; for an entity described by a common noun (such as "the airline"), the probability assigned by the extended grid model is .01. If we suspect (based on MUC6 evidence) that the noun is not coreferent, the probability drops to .006 ("an increase")– if it is a date, it falls even further, to .001. However, given that the entity refers to a person, and some of its mentions are modified, suggesting the article gives a title or description ("Obama's Secretary of State, Hillary Clinton"), the chance that it will be the subject of the next sentence more than triples.

7 Experiments

Table 3 gives results for the extended grid model on the test set. This model is significantly better than the standard grid on discrimination (84% versus 80%) and has a higher mean score on insertion (24% versus 21%)[8].

The best WSJ results in previous work are those of Elsner and Charniak (2008), who combine the entity grid with models based on pronoun coreference and discourse-new NP detection. We report their scores in the table. This comparison is unfair, however, because the improvements from adding non-head nouns improve our baseline grid sufficiently to equal their discrimination result. State-of-the-art results on a different corpus and task were achieved by Soricut and Marcu (2006) using a log-linear mixture of an entity grid, IBM translation models, and a word-correspondence model based on Lapata (2003).

	Disc. Acc	Disc. F	Ins.
Random	50.00	50.00	12.6
Elsner+Charniak	79.6	81.0	23.0
Grid	79.5	80.9	21.4
Extended Grid	**84.0**†	**84.5**	**24.2**
Grid+combo	82.6	84.0	24.3
ExtEGrid+combo	**86.0**†	**86.5**	**26.7**†

Table 3: Extended entity grid and combination model performance on 1004 WSJ test documents. Combination models incorporate pronoun coreference, discourse-new NP detection, and IBM model 1. †indicates an extended model score better than its baseline counterpart at p=.05.

To perform a fair comparison of our extended grid with these model-combining approaches, we train our own combined model incorporating an entity grid, pronouns, discourse-newness and the IBM model. We combine models using a log-linear mixture as in Soricut and Marcu (2006), training the weights to maximize discrimination accuracy.

The second section of Table 3 shows these model combination results. Notably, our extended entity grid on its own is essentially just as good as the combined model, which represents our implementation of the previous state of the art. When we incorporate it into a combination, the performance increase remains, and is significant for both tasks (disc. 86% versus 83%, ins. 27% versus 24%). Though the improvement is not perfectly additive, a good deal of it is retained, demonstrating that our additions to the entity grid are mostly orthogonal to previously described models. These results are the best reported for sentence ordering of English news articles.

8 Conclusion

We improve a widely used model of local discourse coherence. Our extensions to the feature set involve distinguishing simple properties of entities, such as their named entity type, which are also useful in coreference and summarization tasks. Although our method uses coreference information, it does not require coreference resolution to be run on the target documents. Given the popularity of entity grid models for practical applications, we hope our model's improvements will transfer to summarization, generation and readability prediction.

[8] For insertion using the model on its own, the median changes less than the mean, and the change in median score is not significant. However, using the combined model, the change is significant.

Acknowledgements

We are most grateful to Regina Barzilay, Mark Johnson and three anonymous reviewers. This work was funded by a Google Fellowship for Natural Language Processing.

References

Galen Andrew and Jianfeng Gao. 2007. Scalable training of L1-regularized log-linear models. In *ICML '07*.

Regina Barzilay and Mirella Lapata. 2005. Modeling local coherence: an entity-based approach. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*.

Regina Barzilay and Mirella Lapata. 2008. Modeling local coherence: an entity-based approach. *Computational Linguistics*, 34(1):1–34.

Jill Burstein, Joel Tetreault, and Slava Andreyev. 2010. Using entity-based features to model coherence in student essays. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 681–684, Los Angeles, California, June. Association for Computational Linguistics.

Eugene Charniak and Micha Elsner. 2009. EM works for pronoun anaphora resolution. In *Proceedings of EACL*, Athens, Greece.

Eugene Charniak and Mark Johnson. 2005. Coarse-to-fine *n*-best parsing and MaxEnt discriminative reranking. In *Proc. of the 2005 Meeting of the Assoc. for Computational Linguistics (ACL)*, pages 173–180.

Erdong Chen, Benjamin Snyder, and Regina Barzilay. 2007. Incremental text structuring with online hierarchical ranking. In *Proceedings of EMNLP*.

Jackie Chi Kit Cheung and Gerald Penn. 2010. Entity-based local coherence modelling using topological fields. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 186–195, Uppsala, Sweden, July. Association for Computational Linguistics.

Micha Elsner and Eugene Charniak. 2008. Coreference-inspired coherence modeling. In *Proceedings of ACL-08: HLT, Short Papers*, pages 41–44, Columbus, Ohio, June. Association for Computational Linguistics.

Micha Elsner and Eugene Charniak. 2010. The same-head heuristic for coreference. In *Proceedings of ACL 10*, Uppsala, Sweden, July. Association for Computational Linguistics.

Katja Filippova and Michael Strube. 2007. Extending the entity-grid coherence model to semantically related entities. In *Proceedings of the Eleventh European Workshop on Natural Language Generation*, pages 139–142, Saarbrücken, Germany, June. DFKI GmbH. Document D-07-01.

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.

Aria Haghighi and Dan Klein. 2010. Coreference resolution in a modular, entity-centered model. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 385–393, Los Angeles, California, June. Association for Computational Linguistics.

Michael Halliday and Ruqaiya Hasan. 1976. *Cohesion in English*. Longman, London.

Mirella Lapata and Regina Barzilay. 2005. Automatic evaluation of text coherence: Models and representations. In *IJCAI*, pages 1085–1090.

Mirella Lapata. 2003. Probabilistic text structuring: Experiments with sentence ordering. In *Proceedings of the annual meeting of ACL, 2003*.

Neil McIntyre and Mirella Lapata. 2010. Plot induction and evolutionary search for story generation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1562–1572, Uppsala, Sweden, July. Association for Computational Linguistics.

Thomas Morton, Joern Kottmann, Jason Baldridge, and Gann Bierner. 2005. Opennlp: A java-based nlp toolkit. http://opennlp.sourceforge.net.

Ani Nenkova, Advaith Siddharthan, and Kathleen McKeown. 2005. Automatically learning cognitive status for multi-document summarization of newswire. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 241–248, Vancouver, British Columbia, Canada, October. Association for Computational Linguistics.

Emily Pitler, Annie Louis, and Ani Nenkova. 2010. Automatic evaluation of linguistic quality in multi-document summarization. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 544–554, Uppsala, Sweden, July. Association for Computational Linguistics.

Radu Soricut and Daniel Marcu. 2006. Discourse generation using utility-trained coherence models. In *Proceedings of the Association for Computational Linguistics Conference (ACL-2006)*.