Improving Coreference Resolution by Using Conversational Metadata

Xiaoqiang Luo and Radu Florian and Todd Ward
IBM T.J. Watson Research Center
Yorktown Heights, NY 10598
{xiaoluo,raduf,toddward}@us.ibm.com

Abstract

In this paper, we propose the use of metadata contained in documents to improve coreference resolution. Specifically, we quantify the impact of speaker and turn information on the performance of our coreference system, and show that the metadata can be effectively encoded as features of a statistical resolution system, which leads to a statistically significant improvement in performance.

1 Introduction

Coreference resolution aims to find the set of linguistic expressions that refer to a common entity. It is a discourse-level task given that the ambiguity of many referential relationships among linguistic expressions can only be correctly resolved by examining information extracted from the entire document.

In this paper, we focus on exploiting the structural information (e.g., speaker and turn in conversational documents) represented in the metadata of an input document. Such metadata often coincides with the discourse structure, and is presumably useful to coreference resolution. The goal of this study is to quantify the effect metadata. To this end, information contained in metadata is encoded as features in our coreference resolution system, and statistically significant improvement is observed.

The rest of the paper is organized as follows. In Section 2 we describe the data set on which this study is based. In Section 3 we first show how to incorporate information carried by metadata into a statistical coreference resolution system. We also quantify the impact of metadata when they are treated as extraneous data. Results and discussions of the results are also presented in that section.

2 Data Set

This study uses the 2007 ACE data. In the ACE program, a mention is textual reference to an object of interest while the set of mentions in a document referring to the same object is called entity. Each mention is of one of 7 entity types: FAC(cility), GPE (Geo-Political Entity), LOC(ation), ORG(anization), PER(son), VEH(icle), and WEA(pon). Every entity type has a predefined set of subtypes. For example, ORG subtypes include commercial, governmental and educational etc, which reflect different subgroups of organizations. Mentions referring to the same entity share the same type and subtype. A mention can also be assigned with one of 3 mention types: either NAM(e), NOM(inal), or PRO(noun). Accordingly, entities have “levels:” if an entity contains at least one NAM mention, its level is NAM; or if it does not contain any NAM mention, but contains at least one NOM mention, then the entity is of level NOM; if an entity has only PRO mention(s), then its level is PRO. More information about ACE entity annotation can be found in the official annotation guideline (Linguistic Data Consortium, 2008).

The ACE 2007 documents come from a variety of sources, namely newswire, broadcast conversation, broadcast news, Usenet, web log and telephone conversation. Some of them contain rich metadata, as illustrated in the following excerpt of one broadcast conversation document:

```
<DOC>
<DOCID>CNN_CF_20030303.1900.00</DOCID>
<TEXT>
<TURN>
<SPEAKER> Begala </SPEAKER>
Well, we’ll debate that later on in the
show. We’ll have a couple of experts
come out, ...
```

Proceedings of NAACL HLT 2009: Short Papers, pages 201–204,
Boulder, Colorado, June 2009. ©2009 Association for Computational Linguistics

```
</TURN>
<TURN>
<SPEAKER> Novak </SPEAKER>
Paul, as I understand your definition
of a political -- of a professional
politician based on that is somebody
who is elected to public office. ...
</TURN>
...
</TEXT>
</DOC>
```

In this example, `SPEAKER` and `TURN` information are marked by their corresponding SGML tags. Such metadata provides structural information: for instance, the metadata implies that `Begala` is the speaker of the utterance “Well, we’ll debate ..., ” and `Novak` the speaker of the utterance “Paul, as I understand your definition ...” Intuitively, knowing the speakers of the previous and current turn would make it a lot easier to find the right antecedent of pronominal mentions `I` and `your` in the sentence: “Paul, as I understand your definition ...”

Documents in non-conversational genres (e.g. newswire documents) also contain speaker and quotation, which resemble conversational utterance, but they are not annotated. For these documents, we use heuristics (e.g., existence of double or single quote, a short list of communication verb lemmas such as “say,” “tell” and “speak” etc) to determine the speaker of a direct quotation if necessary.

3 Impact of Metadata

In this section we describe how metadata is used to improve our statistical coreference resolution system.

3.1 Resolution System

The coreference system used in our study is a data-driven, machine-learning-based system. Mentions in a document are processed sequentially by mention type: `NAM` mentions are processed first, followed by `NOM` mentions and then `PRO` mentions. The first mention is used to create an initial entity with a deterministic score 1. The second mention can be either linked to the first entity, or used to create a new entity, and the two actions are assigned a score computed from a log linear model. This process is repeated until all mentions in a document are processed. During training time, the process is applied to the training data and training instances (both positive and negative) are generated. At testing time, the same process is applied to an input document and the hypothesis with the highest score is selected as the final coreference result. At the core of the coreference system is a conditional log linear model $P(l|e, m)$ which measures how likely a mention m is or is not coreferential with an existing entity e. The modeling framework provides us with the flexibility to integrate metadata information by encoding it as features.

The coreference resolution system employs a variety of lexical, semantic, distance and syntactic features(Luo et al., 2004; Luo and Zitouni, 2005). The full-blown system achieves an 56.2% ACE-value score on the official 2007 ACE test data, which is about the same as the best-performing system in the Entity Detection and Recognition (EDR) task (NIST, 2007). So we believe that the resolution system is fairly solid.

The aforementioned 56.2% score includes mention detection (i.e., finding mention boundaries and predicting mention attributes) and coreference resolution. Since this study is about coreference resolution only, the subsequent experiments, are thus performed on gold-standard mentions. We split the ACE 2007 data into a training set consisting of 499 documents, and a test set of 100 documents. The training and test split ratio is roughly the same across genres. The performance numbers reported in the subsequent subsections are on the 100-document development test set.

3.2 Metadata Features

For conversational documents with speaker and turn information, we compute a group of binary features for a candidate referent r and the current mention m. Feature values are 1 if the conditions described below hold:

- if r is a speaker, m is a pronominal mention and r utters the sentence containing m.
- if r is a speaker, m is pronoun and r utters the sentence one turn before the one containing m.
- if mention r and mention m are seen in the same turn.
- if mention r and mention m are in two consecutive turns.

Note that the first feature is not subsumed by the third one since a turn may contain multiple sentences. For the same reason, the last feature does not subsume the second one. For the sample document in Section 2, the first feature fires if $r =$ `Novak` and $m =$ `I`; the second features fires if $r =$ `Begala` and m = I; the third feature fires if r = Paul and m = I; and lastly, the fourth feature fires if r = We and m = I. For ACE documents that do not carry turn and speaker information such as newswire, we use heuristic rules to empirically determine the speaker and the corresponding quotations before computing these features.

To test the effect of the feature group, we trained two models: a baseline system without speaker and turn features, and a contrast system by adding the speaker and turn features to the baseline system. The contrast results are tabulated in Table 1. We observe an overall 0.7 point ACE-value improvement. We also compute the ACE-values at document level for the two systems, and a paired Wilcoxon (Wilcoxon, 1945) rank-sum test is conducted, which indicates that the difference between the two systems is statistically significant at level $p \leq 0.002$.

Note that the features often help link pronouns with their antecedents in conversational documents. But ACE-value is a weighted metric which heavily discounts pronominal mentions and entities. We suspect that the effect of speaker and turn information could be larger if we weigh all mention types equally. This is confirmed when we looked at the unweighted B^3 (Bagga and Baldwin, 1998) numbers reported by the official ACE08 scorer (column B^3 in Table 1): the overall B^3 score is improved from 73.8% to 76.4% – a 2.6 point improvement, which is almost 4 times as large as the ACE-value change.

System	ACE-Value	B^3
baseline	78.7	73.8
+ Spkr/Turn	79.4	76.4

Table 1: Coreference performance: baseline vs. system with speaker and turn features.

3.3 Metadata: To Use Or Not to Use?

In the ACE evaluations prior to 2008, mentions inside metadata (such as speaker and poster) are annotated and scored as normal mentions, although such metadata is not part of the actual content of a document. An interesting question is: how large an effect do mentions inside metadata have on the system performance? If metadata are not annotated as mentions, is it still useful to look into them? To answer this question, we remove speaker mentions in conversational documents (i.e., broadcast conversation and telephone conversation) from both the training and test data. Then we train two systems:

- System A: the system totally disregards metadata.
- System B: the system first recovers speaker metadata using a very simple rule: all tokens within the <SPEAKER> tags are treated as one PER mention. This rule recovers most speaker mentions, but it can occasionally result in errors. For instance, the speaker "CNN correspondent John Smith" includes affiliation and profession information and ought to be tagged as three mentions: "CNN" as an ORG(anization) mention, "correspondent" and "John Smith" as two PER mentions. With recovered speaker mentions, we train a model and resolve coreference as normal.

After mentions in the test data are chained in System B, speaker mentions are then removed from system output so that the coreference result is directly comparable with that of System A.

The ACE-value comparison between System A and System B is shown in Table 2. As can be seen, System B works much better than System A, which ignores SPEAKER tags. For telephone conversations (cts), ACE-value improves as much as 4.6 points. A paired Wilcoxon test on document-level ACE-values indicates that the difference is statistically significant at $p < 0.016$.

System	bc	cts
A	75.2	66.8
B	76.6	71.4
Abs. Change	1.4	4.6

Table 2: Metadata improves the ACE-value for broadcast conversation (bc) and telephone conversation (cts) documents.

The reason why metadata helps is that speaker mention can be used to localize the coreference process and therefore improves the performance. For example, in the sentences uttered by "Novak" (cf. the sample document in Section 2), it is intuitively straightforward to link mention I with Novak, and your with Begala – when speaker mentions are made present in the coreference system B. On the other hand, in System A, "I" is likely to be linked with "Paul" because of its proximity of "Paul" in the absence of speaker information.

The result of this experiment suggests that, unsurprisingly, speaker and turn metadata carry structural information helpful for coreference resolution. Even if speaker mentions are not annotated (as in System A), it is still beneficial to make use of it, e.g., by first identifying them automatically as in System B.

4 Related Work

There is a large body of literature for coreference resolution based on machine learning (Kehler, 1997; Soon et al., 2001; Ng and Cardie, 2002; Yang et al., 2008; Luo et al., 2004) approach. Strube and Muller (2003) presented a machine-learning based pronoun resolution system for spoken dialogue (Switchboard corpus). The document genre in their study is similar to the ACE telephony conversation documents, and they did include some dialogue-specific features, such as an anaphora's preference for S, VP or NP, in their system, but they did not use speaker or turn information. Gupta et al. (2007) presents an algorithm disambiguating generic and referential "you."

Cristea et al. (1999) attempted to improve coreference resolution by first analyzing the discourse structure of a document with rhetoric structure theory (RST) (Mann and Thompson, 1987) and then using the resulted discourse structure in coreference resolution. Since obtaining reliably the discourse structure itself is a challenge, they got mixed results compared with a linear structure baseline.

Our work presented in this paper concentrates on the structural information represented in metadata, such as turn or speaker information. Such metadata provides reliable discourse structure, especially for conversational documents, which is proven beneficial for enhancing the performance of our coreference resolution system.

Acknowledgments

This work is partially supported by DARPA GALE program under the contract number HR0011-06-02-0001. We'd also like to thank 3 reviewers for their helpful comments.

References

Amit Bagga and Breck Baldwin. 1998. Algorithms for scoring coreference chains. In *Proceedings of the Linguistic Coreference Workshop at The First International Conference on Language Resources and Evaluation (LREC'98)*, pages 563–566.

Dan Cristea, Nancy lde, Daniel Marcu, Valentin Tablanlivia Polanyi, and Martin van den Berg. 1999. Discourse structure and co-reference: An empirical study. In *Proceedings of ACL Workshop "The Relation of Discourse/Dialogue Structure and Reference"*. Association for Computational Linguistics.

Surabhi Gupta, Matthew Purver, and Dan Jurafsky. 2007. Disambiguating between generic and referential "you" in dialog. In *Proceedings of the 45th ACL(the Demo and Poster Sessions)*, pages 105–108, Prague, Czech Republic, June. Association for Computational Linguistics.

Andrew Kehler. 1997. Probabilistic coreference in information extraction. In *Proc. of EMNLP*.

Linguistic Data Consortium. 2008. ACE (Automatic Content Extraction) English annotation guidelines for entities. http://projects.ldc.upenn.edu/ace/docs/English-Entities-Guidelines_v6.5.pdf.

Xiaoqiang Luo and Imed Zitouni. 2005. Multi-lingual coreference resolution with syntactic features. In *Proc. of Human Language Technology (HLT)/Empirical Methods in Natural Language Processing (EMNLP).*

Xiaoqiang Luo, Abe Ittycheriah, Hongyan Jing, Nanda Kambhatla, and Salim Roukos. 2004. A mention-synchronous coreference resolution algorithm based on the bell tree. In *Proc. of ACL*.

William C. Mann and Sandra A. Thompson. 1987. Rhetorical structure theory: A theory of text organization. Technical Report RS-87-190, USC/Information Sciences Institute.

Vincent Ng and Claire Cardie. 2002. Improving machine learning approaches to coreference resolution. In *Proc. of ACL*, pages 104–111.

NIST. 2007. 2007 automatic content extraction evaluation official results. http://www.nist.gov/speech/tests/ace/2007/doc/ace07_eval_official_results_20070402.html.

Wee Meng Soon, Hwee Tou Ng, and Chung Yong Lim. 2001. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 27(4):521–544.

Michael Strube and Christoph Muller. 2003. A machine learning approach to pronoun resolution in spoken dialogue. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics.*

Frank Wilcoxon. 1945. Individual comparisons by ranking methods. *Biometrics*, I:80–83.

Xiaofeng Yang, Jian Su, Jun Lang, Chew Lim Tan, Ting Liu, and Sheng Li. 2008. An entity-mention model for coreference resolution with inductive logic programming. In *Proceedings of ACL-08: HLT*, pages 843–851, Columbus, Ohio, June. Association for Computational Linguistics.