A Multigraph Model for Coreference Resolution

Sebastian Martschat, Jie Cai, Samuel Broscheit, Éva Mújdricza-Maydt, Michael Strube
Natural Language Processing Group
Heidelberg Institute for Theoretical Studies gGmbH
Heidelberg, Germany
(sebastian.martschat|jie.cai|michael.strube)@h-its.org

Abstract

This paper presents HITS' coreference resolution system that participated in the CoNLL-2012 shared task on multilingual unrestricted coreference resolution. Our system employs a simple multigraph representation of the relation between mentions in a document, where the nodes correspond to mentions and the edges correspond to relations between the mentions. Entities are obtained via greedy clustering. We participated in the closed tasks for English and Chinese. Our system ranked second in the English closed task.

1 Introduction

Coreference resolution is the task of determining which mentions in a text refer to the same entity. This paper describes HITS' system for the CoNLL-2012 Shared Task on multilingual unrestricted coreference resolution, where the goal is to build a system for coreference resolution in an end-to-end multilingual setting (Pradhan et al., 2012). We participated in the closed tasks for English and Chinese and focused on English. Our system ranked second in the English closed task.

Being conceptually similar to and building upon Cai et al. (2011b), our system is based on a directed multigraph representation of a document. A multigraph is a graph where two nodes can be connected by more than one edge. In our model, nodes represent mentions and edges are built from relations between the mentions. The entities to be inferred correspond to clusters in the multigraph.

Our model allows for directly representing any kind of relations between pairs of mentions in a graph structure. Inference over this graph can harness structural properties and the rich set of encoded relations. In order to serve as a basis for further work, the components of our system were designed to work as simple as possible. Therefore our system relies mostly on local information between pairs of mentions.

2 Architecture

Our system is implemented on top of the BART toolkit (Versley et al., 2008). To compute the coreference clusters in a document, we first extract a set of mentions $M = \{m_1, \ldots, m_n\}$ ordered according to their position in the text (Section 2.1). We then build a directed multigraph where the set of nodes is M and edges are induced by relations between mentions (Section 2.4). The relations we use in our system are coreference indicators like string matching or alias (Section 3). For every relation R, we compute a weight w_R using the training data (Section 2.3). We then assign the weight w_R to any edge that is induced by the relation R. Depending on distance and connectivity properties of the graph the weights may change (Section 2.4.1). Given the constructed graph with edge weights, we go through the mentions according to their position in the text and perform greedy clustering (Section 2.6). For Chinese, we employ spectral clustering (Section 2.5) as adopted in Cai et al. (2011b) before the greedy clustering step to reduce the number of candidate antecedents for a mention. The components of our system are described in the following subsections.

Proceedings of the Joint Conference on EMNLP and CoNLL: Shared Task, pages 100–106,
Jeju Island, Korea, July 13, 2012.

2.1 Mention Extraction

Noun phrases are extracted from the provided parse and named entity annotation layers. For embedded mentions with the same head, we only keep the mention with the largest span.

2.1.1 English

For English we identify eight different mention types: common noun, proper noun, personal pronoun, demonstrative pronoun, possessive pronoun, coordinated noun phrase, quantifying noun phrase (*some of ...*, *17 of ...*) and quantified noun phrase (*the armed men* in *one of the armed men*). The head for a common noun or a quantified noun is computed using the SemanticHeadFinder from the Stanford Parser[1]. The head for a proper noun starts at the first token tagged as a noun until a punctuation, preposition or subclause is encountered. Coordinations have the CC tagged token as head and quantifying noun phrases have the quantifier as head.

In a postprocessing step we filter out adjectival use of nations and named entities with semantic class *Money*, *Percent* or *Cardinal*. We discard mentions whose head is embedded in another mention's head. Pleonastic pronouns are identified and discarded via a modified version of the patterns used by Lee et al. (2011).

2.1.2 Chinese

For Chinese we detect four mention types: common noun, proper noun, pronoun and coordination. The head detection for Chinese is provided by the SunJurafskyChineseHeadFinder from the Standford Parser, except for proper nouns whose head is set to the mention's rightmost token.

The remaining processing is similar to the mention detection for English.

2.2 Preprocessing

We extract the information in the provided annotation layers and transform the predicted constituent parse trees into dependency parse trees. We work with two different dependency representations, one obtained via the converter implemented in Stanford's NLP suite[2], the other using LTH's constituent-to-dependency conversion tool[3]. For pronouns, we determine number and gender using tables containing a mapping of pronouns to their gender and number.

2.2.1 English

For English, number and gender for common nouns are computed via a comparison of head lemma to head and using the number and gender data of Bergsma and Lin (2006). Quantified noun phrases are always plural. We compute semantic classes via a WordNet (Fellbaum, 1998) lookup.

2.2.2 Chinese

For Chinese, we simply determine number and gender by searching for the corresponding designators, since plural mentions mostly end with 们, while 先生 (*sir*) and 女士 (*lady*) often suggest gender information. To identify demonstrative and definite noun phrases, we check whether they start with a definite/demonstrative indicator (e.g. 这 (*this*) or 那 (*that*)). We use lists of named entities extracted from the training data to determine named entities and their semantic class in development and testing data.

2.3 Computing Weights for Relations

We compute weights for relations using simple descriptive statistics on training documents. Since this is a robust approach to learning weights for the type of graph model we employ (Cai et al., 2011b; Cai et al., 2011a), we use only a fraction of the available training data. We took a random subset consisting of around 20% for English and 15% for Chinese of the training data. For every document in this set and every relation R, we go through the set M of extracted mentions and compute for every pair (m_i, m_j) with $i > j$ whether R holds for this pair. The weight w_R for R is then the number of coreferent pairs with R divided by the number of all pairs with R.

2.4 Graph Construction

The set of relations we employ consists of two subsets: negative relations R_- which enforce that no

[1] http://nlp.stanford.edu/software/lex-parser.shtml

[2] http://nlp.stanford.edu/software/stanford-dependencies.shtml

[3] http://nlp.cs.lth.se/software/treebank_converter/

edge is built between two mentions, and positive relations R_+ that induce edges. Again, we go through M in a left-to-right fashion. If for two mentions m_i, m_j with $i > j$ a negative relation R_- holds, no edge between m_i and m_j can be built. Otherwise we add an edge from m_i to m_j for every positive relation R_+ such that $R_+(m_i, m_j)$ is true. The structure obtained by this construction is a directed multigraph.

We handle copula relations similar to Lee et al. (2011): if m_i is *this* and the pair (m_i, m_j) is in a copula relation (like *This is the World*), we remove m_j and replace m_j in all edges involving it by m_i. For Chinese, we handle "role appositives" as introduced by Haghighi and Klein (2009) analogously.

2.4.1 Assigning Weights to Edges

Initially, any edge (m_i, m_j) induced by the relation R has the weight w_R computed as described in Section 2.3. If R is a transitive relation, we divide the weight by the number of mentions connected by this relation. This corresponds to the way edge weights are assigned during the spectral embedding in Cai et al. (2011b). If R is a relation sensitive to distance like compatibility between a common/proper noun and a pronoun, the weight is altered according to the distance between m_i and m_j.

2.4.2 An Example

We demonstrate the graph construction by a simple example. Consider a document consisting of the following three sentences.

> *Barack Obama and Nicolas Sarkozy met in Toronto yesterday. They discussed the financial crisis. Sarkozy left today.*

Let us assume that our system identifies *Barack Obama* (m_1), *Nicolas Sarkozy* (m_2), *Barack Obama and Nicolas Sarkozy* (m_3), *They* (m_4) and *Sarkozy* (m_5) as mentions. We consider these mentions and the relations N_Number, P_Nprn_Prn, P_Alias and P_Subject described in Section 3. The graph constructed according to the algorithm described in this section is displayed in Figure 1.

Observe the effect of the negative relation N_Number: while P_Nprn_Prn holds for the pair *Barack Obama* (m_1) and *They* (m_4), the mentions do not agree in number. Hence N_Number holds for this pair and no edge from m_4 to m_1 can be built.

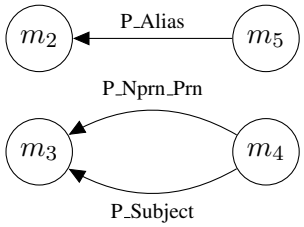

Figure 1: An example graph. Nodes represent mentions, edges are induced by relations between the mentions.

2.5 Spectral Clustering

For Chinese we apply spectral clustering before the final greedy clustering phase. In order to be able to apply spectral clustering, we make the graph undirected and merge parallel edges into one edge, summing up all weights. Due to the way edge weights are computed, the resulting undirected simple graph corresponds to the graph Cai et al. (2011b) use as input to the spectral clustering algorithm. Spectral clustering is now performed as in Cai et al. (2011b).

2.6 Greedy Clustering

To describe our clustering algorithm, we use some additional terminology: if there exists an edge from m to n we say that m is a *parent* of n and that n is a *child* of m.

In the last step, we go through the mentions from left to right. Let m_i be the mention in focus. For English, we consider all children of m_i as possible antecedents. For Chinese we restrict the possible antecedents to all children that are in the same cluster obtained by spectral clustering.

If m_i is a pronoun, we determine m_j such that the sum over all weights of edges from m_i to m_j is maximized. We then assign m_i and m_j to the same entity. In English, if m_i is a parent of a noun phrase m that embeds m_j, we instead assign m_i and m to the same entity.

For Chinese, all other noun phrases are assigned to the same entity as all their children in the cluster obtained by spectral clustering. For English, we are more restrictive: definites and demonstratives are assigned to the same cluster as their closest (according to the position of the mentions in the text) child.

Negative relations may also be applied as constraints in this phase. Before assigning m_i to the same entity as a set of mentions C, we check for

every $m \in C$ and every negative relation R_- that we want to incorporate as a constraint whether $R_-(m_i, m)$ holds. If yes, we do not assign m_i to the same entity as the mentions in C.

2.7 Complexity

Our algorithms for weight computation, graph construction and greedy clustering look at all pairs of mentions in a document and perform simple calculations, which leads to a time complexity of $O\left(n^2\right)$ per document, where n is the number of mentions in a document. When performing spectral clustering, this increases to $O\left(n^3\right)$. Since we deal with at most a few hundred mentions per document, the cubic running time is not an issue.

3 Relations

In our system relations serve as templates for building or disallowing edges between mentions. We distinguish between positive and negative relations: negative relations disallow edges between mentions, positive relations build edges between mentions. Negative relations can also be used as constraints during clustering, while positive relations may also be applied as "weak" relations: in this case, we only add the induced edge when the two mentions under consideration are already included in the graph after considering all the non-weak relations.

Most of the relations presented here were already used in our system for last year's shared task (Cai et al., 2011b). The set of relations was enriched mainly to resolve pronouns in dialogue and to resolve pronouns that do not carry much information by themselves like *it* and *they*.

3.1 Negative Relations

(1) N_Gender, (2) N_Number: Two mentions do not agree in gender or number.

(3) N_SemanticClass: Two mentions do not agree in semantic class (only the *Object*, *Date* and *Person* top categories derived from WordNet (Fellbaum, 1998) are used).

(4) N_ItDist: The anaphor is *it* or *they* and the sentence distance to the antecedent is larger than one.

(5) N_BarePlural: Two mentions that are both bare plurals.

(6) N_Speaker12Prn: Two first person pronouns or two second person pronouns with different speakers, or one first person pronoun and one second person pronoun with the same speaker.

(7) N_DSprn: Two first person pronouns in direct speech assigned to different speakers.

(8) N_ContraSubjObj: Two mentions are in the subject and object positions of the same verb, and the anaphor is a non-possessive pronoun.

(9) N_Mod: Two mentions have the same syntactic heads, and the anaphor has a pre- or post-modifier which does not occur in the antecedent and does not contradict the antecedent.

(10) N_Embedding: Two mentions where one embeds the other, which is not a reflexive or possessive pronoun.

(11) N_2PrnNonSpeech: Two second person pronouns without speaker information and not in direct speech.

3.2 Positive Relations

(12) P_StrMatch_Npron, (13) P_StrMatch_Pron: After discarding stop words, if the strings of mentions completely match and are not pronouns, the relation P_StrMatch_Npron holds. When the matched mentions are pronouns, P_StrMatch_Pron holds.

(14) P_HeadMatch: If the syntactic heads of mentions match.

(15) P_Nprn_Prn: If the antecedent is not a pronoun and the anaphor is a pronoun. This relation is restricted to a sentence distance of 1.

(16) P_Alias: If mentions are aliases of each other (i.e. proper names with partial match, full names and acronyms, etc.).

(17) P_Speaker12Prn: If the speaker of the second person pronoun is talking to the speaker of the first person pronoun. The mentions contain only first or second person pronouns.

(18) P_DSPrn: If one mention is subject of a *speak* verb, and the other mention is a first person pronoun within the corresponding direct speech.

(19) P_ReflPrn: If the anaphor is a reflexive pronoun, and the antecedent is the subject of the sentence.

(20) P_PossPrn: If the anaphor is a possessive pronoun, and the antecedent is the subject of the sentence or the subclause.

(21) P_GPEIsA: If the antecedent is a Named Entity of *GPE* entity type and the anaphor is a definite expression of the same type.

(22) P_PossPrnEmbedding: If the anaphor is a possessive pronoun and is embedded in the antecedent.

(23) P_VerbAgree: If the anaphor is a pronoun and has the same predicate as the antecedent.

(24) P_Subject & (25) P_Object: If both mentions are subjects/objects (applies only if the anaphor is *it* or *they*).

(26) P_SemClassPrn: If the anaphor is a pronoun, the antecedent is not a pronoun, and both have semantic class *Person*.

For English, we used all relations except for (21) and (26). Relations (1), (2) and (10) were incorporated as constraints during greedy clustering. For Chinese, we used relations (1) – (6), (12) – (15), (21) and (26). (26) was incorporated as a weak relation.

4 Results

We submitted to the closed tasks for English and Chinese. The results on the English development set and testing set are displayed in Table 1 and Table 2 respectively. To indicate the progress we achieved within one year, Table 3 shows the performance of our system on the CoNLL '11 development data set compared to last year's results (Cai et al., 2011b). The *Overall* number is the average of MUC, B^3 and CEAF (E), MD is the mention detection score. Overall, we gained over 5% F1 some of which can be attributed to improved mention detection.

Metric	R	P	F1
MD	73.96	75.69	74.81
MUC	64.93	68.69	66.76
B^3	68.42	75.77	71.91
CEAF (M)	61.23	61.23	61.23
CEAF (E)	49.61	45.60	47.52
BLANC	77.81	80.75	79.19
Overall			62.06

Table 1: Results on the English CoNLL '12 development set

Metric	R	P	F1
MD	74.23	76.10	75.15
MUC	65.21	68.83	66.97
B^3	66.50	74.69	70.36
CEAF (M)	59.61	59.61	59.61
CEAF (E)	48.64	44.72	46.60
BLANC	73.29	78.94	75.73
Overall			61.31

Table 2: Results on the English CoNLL '12 testing set

Metric	R	P	F1	2011 F1
MD	70.84	73.08	71.94	66.28
MUC	60.80	65.09	62.87	55.19
B^3	68.37	75.89	71.94	68.52
CEAF (M)	60.42	60.42	60.42	54.44
CEAF (E)	50.40	46.11	48.16	43.19
BLANC	75.44	79.26	77.19	72.13
Overall			60.99	55.63

Table 3: Results on the English CoNLL '11 development set compared to Cai et al. (2011b)

Table 4 and Table 5 display our results on Chinese development data and testing data respectively.

Metric	R	P	F1
MD	52.45	71.50	60.51
MUC	45.90	67.07	54.50
B^3	58.94	84.26	69.36
CEAF (M)	53.60	53.60	53.60
CEAF (E)	50.73	34.24	40.89
BLANC	66.17	83.11	71.45
Overall			54.92

Table 4: Results on the Chinese CoNLL '12 development set

Metric	R	P	F1
MD	48.49	74.02	58.60
MUC	42.71	67.80	52.41
B^3	55.37	85.24	67.13
CEAF (M)	51.30	51.30	51.30
CEAF (E)	51.81	32.46	39.92
BLANC	63.96	82.81	69.18
Overall			53.15

Table 5: Results on the Chinese CoNLL '12 testing set

Because none of our team members has knowledge of the Arabic language we did not attempt to

run our system on the Arabic datasets and therefore our official score for this language is considered to be 0. The combined official score of our submission is $(0.0 + 53.15 + 61.31)/3 = 38.15$. In the closed task our system was the second best performing system for English and the eighth best performing system for Chinese.

5 Error analysis

We did not attempt to resolve event coreference and did not incorporate world knowledge which is responsible for many recall errors our system makes.

Since we use a simple greedy strategy for clustering that goes through the mentions left-to-right, errors in clustering propagate, which gives rise to cluster-level inconsistencies. We observed a drop in performance when using more negative relations as constraints. A more sophisticated clustering strategy that allows a more refined use of constraints is needed.

5.1 English

Our detection of copula and appositive relations is quite inaccurate, which is why we limit the incorporation of copulas to cases where the antecedent is *this* and left appositives out.

We aim for high precision regarding the usage of the negative relation N_Modifier. This leads to some loss in recall. For example, our system does not assign *the just-completed Paralympics* and *the 12-day Paralympics* to the same entity. Such cases require a more involved reasoning scheme to decide whether the modifiers are actually contradicting each other.

Non-referring pronouns constitute another source of errors. While we improved detection of pleonastic *it* compared to last year's system, a lot of them are not filtered out. Our system also does not distinguish well between generic and non-generic uses of *you* and *we*, which hurts precision.

5.2 Chinese

Since each Chinese character carries its own meaning, there are multiple ways to express the same entity by combining different characters into a word. Both syntactic heads and modifiers can be replaced by similar words or by abbreviated versions. From 外省人 (*outside people*) to 外省 族群 (*outside ethnic group*) the head is replaced, while from 戴安娜 (*Diana*) to 美丽 迷人 的 戴妃 (*charming Di Princess*) the name is abbreviated.

Modifier replacement is more difficult to cope with, our system does not recognize that 重新 计票 作业 (*starting-over counting-votes job*) and 验票 作业 (*verifying-votes job*) are coreferent. It is also not trivial to separate characters from words (e.g. by separating 计 and 票) to resolve such cases, since it will introduce too much noise as a consequence. In order to tackle this problem, a smart scheme to propagate similarities from partial words to the entire mentions and a knowledge base upon which reliable similarities can be retrieved are necessary.

In contrast to English there is no strict enforcement of using definite noun phrases when referring to an antecedent in Chinese. Both 这次演说 (*the talk*) and 演说 (*talk*) can corefer with the antecedent 克林顿 在 河内 大选 的 演说 (*Clinton's talk during Hanoi election*). This makes it very difficult to distinguish generic expressions from referential ones. In the submitted version of our system, we simply ignore the nominal anaphors which do not start with definite articles or demonstratives.

6 Conclusions

In this paper we presented a graph-based model for coreference resolution. It captures pairwise relations between mentions via edges induced by relations. Entities are obtained by graph clustering. Discriminative information can be incorporated as negative relations or as constraints during clustering.

We described our system's architecture and the relations it employs, highlighting differences and similarities to our system from last year's shared task.

Designed to work as a basis for further work, our system works mainly by exploring the relationship between pairs of mentions. Due to its modular architecture, our system can be extended by components taking global information into account, for example for weight learning or clustering.

We focused on the closed task for English in which our system achieved competitive performance, being ranked second out of 15 participants.

Acknowledgments. This work has been funded by the Klaus Tschira Foundation, Heidelberg, Germany. The first and the second authors have been supported by a HITS PhD. scholarship.

References

Shane Bergsma and Dekang Lin. 2006. Bootstrapping path-based pronoun resolution. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics,* Sydney, Australia, 17–21 July 2006, pages 33–40.

Jie Cai, Éva Mújdricza-Maydt, Yufang Hou, and Michael Strube. 2011a. Weakly supervised graph-based coreference resolution for clinical data. In *Proceedings of the 5th i2b2 Shared Tasks and Workshop on Challenges in Natural Language Processing for Clinical Data,* Washington, D.C., 20-21 October 2011. To appear.

Jie Cai, Éva Mújdricza-Maydt, and Michael Strube. 2011b. Unrestricted coreference resolution via global hypergraph partitioning. In *Proceedings of the Shared Task of the 15th Conference on Computational Natural Language Learning,* Portland, Oreg., 23–24 June 2011, pages 56–60.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, Mass.

Aria Haghighi and Dan Klein. 2009. Simple coreference resolution with rich syntactic and semantic features. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing,* Singapore, 6–7 August 2009, pages 1152–1161.

Heeyoung Lee, Yves Peirsman, Angel Chang, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2011. Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 shared task. In *Proceedings of the Shared Task of the 15th Conference on Computational Natural Language Learning,* Portland, Oreg., 23–24 June 2011, pages 28–34.

Sameer Pradhan, Alessandro Moschitti, and Nianwen Xue. 2012. CoNLL-2012 Shared Task: Modeling multilingual unrestricted coreference in OntoNotes. In *Proceedings of the Shared Task of the 16th Conference on Computational Natural Language Learning,* Jeju Island, Korea, 12–14 July 2012. This volume.

Yannick Versley, Simone Paolo Ponzetto, Massimo Poesio, Vladimir Eidelman, Alan Jern, Jason Smith, Xiaofeng Yang, and Alessandro Moschitti. 2008. BART: A modular toolkit for coreference resolution. In *Companion Volume to the Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics,* Columbus, Ohio, 15–20 June 2008, pages 9–12.