BART goes multilingual: The UniTN / Essex submission to the CoNLL-2012 Shared Task

Olga Uryupina‡ **Alessandro Moschitti**‡ **Massimo Poesio**‡†
‡University of Trento
† University of Essex
uryupina@gmail.com, moschitti@disi.unitn.it, massimo.poesio@unitn.it

Abstract

This paper describes the UniTN/Essex submission to the CoNLL-2012 Shared Task on the Multilingual Coreference Resolution. We have extended our CoNLL-2011 submission, based on BART, to cover two additional languages, Arabic and Chinese. This paper focuses on adapting BART to new languages, discussing the problems we have encountered and the solutions adopted. In particular, we propose a novel entity-mention detection algorithm that might help identify nominal mentions in an unknown language. We also discuss the impact of basic linguistic information on the overall performance level of our coreference resolution system.

1 Introduction

A number of high-performance coreference resolution (CR) systems have been created for English in the past decades, implementing both rule-based and statistical approaches. For other languages, however, the situation is far less optimistic. For Romance and German languages, several systems have been developed and evaluated, in particular, at the SemEval-2010 track 1 on Multilingual Coreference Resolution (Recasens et al., 2010). For other languages, individual approaches have been proposed, covering specific subparts of the task, most commonly pronominal anaphors (cf., for example, (Iida and Poesio, 2011; Arregi et al., 2010) and many others).

Two new languages, Arabic and Chinese, have been proposed for the CoNLL-2012 shared task (Pradhan et al., 2012). They present a challenging problem: the systems are required to provide entity mention detection (EMD) and design a proper coreference resolver for both languages. At UniTN/Essex, we have focused on these parts of the task, relying on a modified version of our last-year submission for English.

Most state-of-the-art full-scale coreference resolution systems rely on hand-written rules for the mention detection subtask.[1] For English, such rules may vary from corpus to corpus, reflecting specifics of particular guidelines (e.g. whether nominal premodifiers can be mentions, as in MUC, or not, as in most other corpora). However, for each corpus, such heuristics can be adjusted in a straightforward way. Creating a robust rule-based EMD module for a new language, on the contrary, is a challenging issue that requires substantial linguistic knowledge.

In this paper, we advocate a novel approach, recasting parse-based EMD as a statistical problem. We consider a node-filtering model that does not rely on any linguistic expertise in a given language. Instead, we use tree kernels (Moschitti, 2008; Moschitti, 2006) to induce a classifier for mention NP-nodes automatically from the data.

Another issue to be solved when designing a coreference resolution system for a new language is a possible lack of relevant linguistic information. Most state-of-the-art CR algorithms rely on relatively advanced linguistic representations of mentions. This can be seen as a remarkable shift

[1] Statistical EMD approaches have been proved useful for ACE-style coreference resolution, where mentions are basic units belonging to a restricted set of semantic types.

from knowledge-lean approaches of the late nineties (Harabagiu and Maiorano, 1999). In fact, modern systems try to account for complex coreference links by incorporating lexicographic and world knowledge, for example, using WordNet (Harabagiu et al., 2001; Huang et al., 2009) or Wikipedia (Ponzetto and Strube, 2006). For languages other than English, however, even the most basic properties of mentions can be intrinsically difficult to extract. For example, Baran and Xue (2011) have shown that a complex algorithm is needed to identify the `number` property of Chinese nouns.

Both Arabic and Chinese have long linguistic traditions and therefore most grammar studies rely on terminology that can be very confusing for an outsider. For example, several works on Arabic (Hoyt, 2008) mention that nouns can be made definite with the suffix "Al-", but this is not a semantic, but syntactic definiteness. Without any experience in Arabic, one can hardly decide how such "syntactic definiteness" might affect coreference.

In the present study, we have used the information provided by the CoNLL organizers to try and extract at least some linguistic properties of mentions for Arabic and Chinese. We have run several experiments, evaluating the impact of such very basic knowledge on the performance level of a coreference resolution system.

The rest of the paper is organized as follows. In the next section we briefly describe the general architecture and the system for English, focusing on the adjustments made after the last year competition. Section 3 is devoted to new languages: we first discuss our EMD module and then describe the procedures for extracting linguistic knowledge. Section 4 discusses the impact of our solutions to the performance level of a coreference resolver. The official evaluation results are presented in Section 5.

2 BART

Our CoNLL submission is based on BART (Versley et al., 2008). BART is a modular toolkit for coreference resolution that supports state-of-the-art statistical approaches to the task and enables efficient feature engineering. BART has originally been created and tested for English, but its flexible modular architecture ensures its portability to other languages and domains.

The BART toolkit has five main components: preprocessing pipeline, mention factory, feature extraction module, decoder and encoder. In addition, an independent *LanguagePlugin* module handles all the language specific information and is accessible from any component.

The architecture is shown in Figure 1. Each module can be accessed independently and thus adjusted to leverage the system's performance on a particular language or domain.

The preprocessing pipeline converts an input document into a set of linguistic layers, represented as separate XML files. The mention factory uses these layers to extract mentions and assign their basic properties (number, gender etc). The feature extraction module describes pairs of mentions $\{M_i, M_j\}$, $i < j$ as a set of features. At the moment we have around 45 different feature extractors, encoding surface similarity, morphological, syntactic, semantic and discourse information. Note that no language-specific information is encoded in the extractors explicitly: a language-independent representation, provided by the Language Plugin, is used to compute feature values. For CoNLL-2012, we have created two additional features: `lemmata-match` (similar to string match, but uses lemmata instead of tokens) and `number-agreement-du` (similar to commonly used number agreement features, but supports dual number).

The encoder generates training examples through a process of sample selection and learns a pairwise classifier. Finally, the decoder generates testing examples through a (possibly distinct) process of sample selection, runs the classifier and partitions the mentions into coreference chains.

2.1 Coreference resolution in English

The English track at CoNLL-2012 can be considered an extension of the last year's CoNLL task. New data have been added to the corpus, including two additional domains, but the annotation guidelines remain the same.

We have therefore mainly relied on the CoNLL-2011 version of our system (Uryupina et al., 2011) for the current submission, providing only minor adjustments. Thus, we have modified our preprocess-

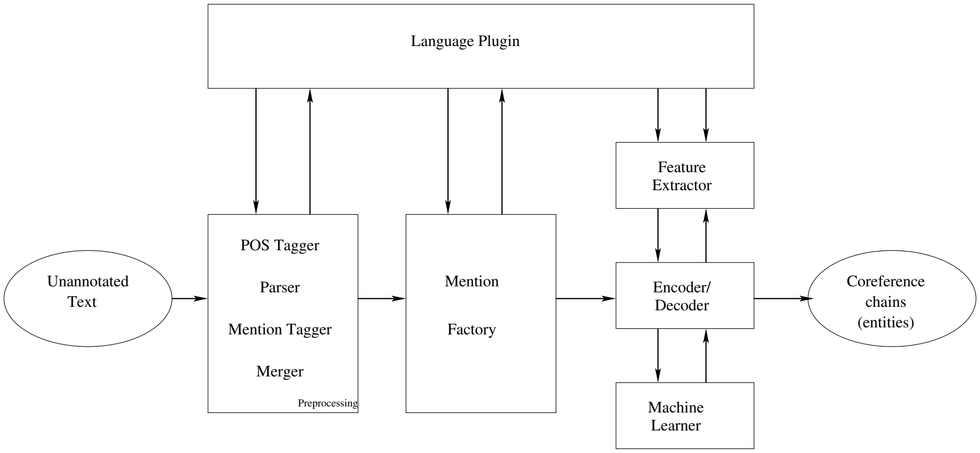

Figure 1: BART architecture

ing pipeline to operate on the OntoNotes NE-types, mapping them into MUC types required by BART. This allows us to participate in the closed track, as no external material is used any longer.

Since last year, we have continued with our experiments on multi-objective optimization, proposed in our CoNLL-2011 paper (Uryupina et al., 2011). We have extended the scope of our work to cover different machine learning algorithms and their parameters (Saha et al., 2011). For CoNLL-2012, we have re-tested all the solutions of our optimization experiments, picking the one with the highest score on the current development set.

Finally, our recent experiments on domain selection (Uryupina and Poesio, 2012) suggest that, at least for some subparts of OntoNotes, a system might benefit from training a domain-specific model. We have tested this hypothesis on the CoNLL-2012 data and have consequently trained domain-specific classifiers for the nw and bc domains.

3 Coreference resolution in Arabic and Chinese

We have addressed two main issues when developing our coreference resolvers for Arabic and Chinese: mention detection and extraction of relevant linguistic properties of our mentions.

3.1 Mention detection

Mention detection is rarely considered to be a separate task. Only very few studies on coreference resolution report on their EMD techniques. Existing corpora of coreference follow different approaches to mention annotation: this includes defining mention boundaries (basic vs. maximal NPs), alignment procedures (strict vs. relaxed with manually annotated minimal spans vs. relaxed with automatically extracted heads), the position on singleton and/or non-referential mentions (annotated vs. not).

The CoNLL-2011/2012 guidelines take a very strict view on mention boundaries: only the maximal spans are annotated and no approximate matching is allowed. Moreover, the singleton mentions (i.e. not participating in coreference relations) are not marked. This makes the mention detection task for OntoNotes extremely challenging, especially for the two new languages: on the one hand, one has to provide exact boundaries; on the other hand, it is hard to learn such information explicitly, as not all the candidate mentions are annotated.

Most CoNLL-2011 systems relied on hand-written rules for the mention detection subtask. This was mainly possible due to the existence of well-studied and thoroughly documented head-detection rules for English, available as a description for reimplementing (Collins, 1999) or as a downloadable package. Consider the following example:

(1) ..((the rising price)$_{NP_2}$ of (food)$_{NP_3}$)$_{NP_1}$..

In this fragment, three nominal phrases can be identified, with the first one ("the rising price of food") spanning over the two others ("the rising price") and ("food"). According to the OntoNotes annotation guidelines, the second noun phrase cannot be a mention, because it is embedded in an upper NP and they share the same head noun. The third noun phrase, on the contrary, could be a mention—even though it's embedded in another NP, their heads are different. Most CoNLL-2011 participants used as a backbone a heuristic discarding embedded noun phrases.

For less-known languages, however, this heuristic is only applicable as long as we can compute an NP's head reliably. Otherwise it's hard to distinguish between candidate mentions similar to NP_1 and to NP_2 in the example above.

A set of more refined heuristics is typically applied to discard or add some specific types of mentions. For example, several studies (Bergsma and Yarowsky, 2011) have addressed the issue of detecting expletive pronouns in English. Again, in the absence of linguistic expertise, one can hardly engineer such heuristics for a new language manually.

We have investigated the possibility of learning mention boundaries automatically from the OntoNotes data. We recast the problem as an NP-node filtering task: we analyze automatically computed parse trees and consider all the NP-nodes to be candidate instances to learn a classifier of correct vs. incorrect mention nodes. Clearly, this approach cannot account for mentions that do not correspond to NP-nodes. However, as Table 1 shows, around 85-89% of all the mentions, both for Arabic and Chinese, are NP-nodes.

	train		development	
	NP-nodes	%	NP-nodes	%
Arabic	24068	87.23	2916	87.91
Chinese	88523	85.96	12572	88.52

Table 1: NP-nodes in OntoNotes for Arabic and Chinese: total numbers and percentage of mentions.

We use tree kernels (Moschitti, 2008; Moschitti, 2006) to induce a classifier that labels an NP node and a part of the parse tree that surrounds it as ±mention. Two integer parameters control the selection of the relevant part of the parse tree, allowing for pruning the nodes that are far above or far below the node of interest.

Our classifier is supposed to decide whether an NP-node is a mention of a real-world object. Such mentions, however, are annotated in OntoNotes as positive instances only when they corefer with some other mentions. The classifier works as a preprocessor for a CR system and therefore has no information that would allow it to discriminate between singleton vs. non-singleton mentions. One can investigate possibilities for joint EMD and CR to alleviate the problem. We have adopted a simpler solution: we tune a parameter (cost factor) that controls the precision/recall trade-off to bias the classifier strongly towards recall.

We use a small subset (1-5%) of the training data to train the EMD classifier. We tune the EMD parameters to optimize the overall performance: we run the classifier to extract mentions for the whole training and development sets, run the coreference resolver and record the obtained result (CoNLL score). The whole set of parameters to be tuned comprise: the size of the training set for EMD, the precision-recall trade-off, and two pruning thresholds.

3.2 Extracting linguistic properties

All the features implemented in BART use some kind of linguistic information from the mentions. For example, the `number-agreement` feature first extracts the `number` properties of individual mentions. For a language supported by BART, such properties are computed by the MentionFactory. For a new language, they should be provided as a part of the mention representation computed by some external preprocessing facilities. The only obligatory mention property is its span— the sequence of relevant token ids—all the properties discussed below are optional.

The following properties have been extracted for new languages directly from the CoNLL table:

- sentence id
- sequence of lemmata
- speaker (Chinese only)

Coordinations have been determined by analyzing the sequence of PoS tags: any span containing a coordinate conjunction is a coordination. They are always considered plural and unspecified for gender, their heads correspond to their entire spans.

For non-coordinate NPs, we extract the head nouns using simple heuristics. In Arabic, the first noun in a sequence is a head. In Chinese, the last one is a head. If no head can be found through this heuristic, we try the same method, but allow for pronouns to be heads, and, as a default, consider the whole span to be the head.

Depending on the PoS tag of the head noun, we classify a mention as an NE, a pronoun or a nominal (default). For named entities, no further mention properties have been extracted.

We have compiled lists of pronouns for both Arabic and Chinese from the training and development data. For Arabic, we use gold PoS tags to classify pronouns into subtypes, person, number and gender. For Chinese, no such information is available, so we have consulted several grammar sketches and lists of pronouns on the web. We do not encode clusivity[2] and honorifics.[3]

For Arabic, we extract the gender and number properties of nominals in the following way. First, we have processed the gold PoS tags to create a list of number and gender affixes. We compute the properties of our mentions by analyzing the affixes of their heads. In a number of constructions, however, the gender is not marked explicitly, so we have compiled a gender dictionary for Arabic lemmata on the training and development data. If the gender cannot be computed from affixes, we look it up in the dictionary.

Finally, we have made an attempt at computing the definiteness of nominal expressions. For Arabic, we consider as definites all mentions with definite head nouns (prefixed with "Al") and all the idafa constructs with a definite modifier.[4] We could not compute definiteness for Chinese reliably.

4 Evaluating the impact of kernel-based mention detection and basic linguistic knowledge

To adopt our system to new languages, we have focused on two main issues: EMD and extraction of linguistic properties. In this section we discuss the impact of each factor on the overall performance. Table 2 summarizes our evaluation experiments. All the figures reported in this section are CoNLL scores (averages of MUC, B^3 and $CEAF_e$) obtained on the development data.

To evaluate the impact of our kernel-based EMD (TKEMD), we compare its performance against two baselines. The lower bound, "allnp", considers all the NP-nodes in a parse tree to be candidate mentions. The upper bound, "goldnp" only considers gold NP-nodes to be candidate mentions. Note that the upper bound does not include mentions that do not correspond to NP-nodes at all (around 12% of all the mentions in the development data, cf. Table 1 above).

We have created three versions of our coreference resolver, using different amounts of linguistic knowledge. The baseline system (Table 2, first column) relies only on mention spans. The system itself is a reimplementation of Soon et al. (2001), but, clearly, only the string-matching feature can be computed without specifying mention properties.

A more advanced version of the system (second column) uses the same model and the same feature set, but relies on mention properties, extracted as described in Section 3.2 above. The final version (third column) makes use of all the features implemented in BART. We run a greedy feature selection algorithm, starting from the string matching and adding features one by one, until the performance stops increasing.

For Chinese, our EMD approach has proved to be useful, bringing around 1.5-2% improvement over the "allnp" baseline for all the versions of the coreference resolver. The module for extracting mention properties has only brought a moderate improvement. This is not surprising, as we have not been able to extract many relevant linguistic properties, especially for nominals. We believe that an improvement can be achieved on the Chinese data by incorporating more linguistic information.

[2] In some dialects of Chinese, a distinction is made between the first person plural inclusive ("you and me") and the first person exclusive ("me and somebody else") pronouns.

[3] In Chinese, different pronouns should be used addressing different persons, reflecting the relative social status of the speaker and the listener.

[4] Idafa-constructs are syntactic structures, conveying, very roughly speaking, genitive semantics, commonly used in Arabic. Their accurate analysis requires some language-specific processing.

	baseline	+linguistics	+linguistics
			+features
Arabic			
allnp	45.47	46.15	46.32
TKEMD	46.98	47.44	49.07
goldnp	51.08	63.27	64.55
Chinese			
allnp	50.72	51.04	51.40
TKEMD	53.10	53.33	53.53
goldnp	57.78	57.30	57.98

Table 2: Evaluating the impact of EMD and linguistic knowledge: CoNLL F-score.

For Arabic, the linguistic properties could potentially be very helpful: on gold NPs, our linguistically rich system outperforms its knowledge-lean counterpart by 13 percentage points. Unfortunately, this improvement is mirrored only partially on the fully automatically acquired mentions.

5 Official results

Table 3 shows the official results obtained by our system at the CoNLL-2012 competition.

Metric	Recall	Precision	F-score
English			
MUC	61.00	60.78	60.89
BCUBED	63.59	68.48	65.95
CEAF (M)	52.44	52.44	52.44
CEAF (E)	41.42	41.64	41.53
BLANC	67.40	72.83	69.65
Arabic			
MUC	41.33	41.66	41.49
BCUBED	65.77	69.23	67.46
CEAF (M)	50.82	50.82	50.82
CEAF (E)	42.43	42.13	42.28
BLANC	65.58	70.56	67.69
Chinese			
MUC	45.62	63.13	52.97
BCUBED	59.17	80.78	68.31
CEAF (M)	52.40	52.40	52.40
CEAF (E)	48.47	34.52	40.32
BLANC	68.72	80.76	73.11

Table 3: BART performance at CoNLL-2012: official results on the test set.

6 Conclusion

In this paper we have discussed our experiments on adapting BART to two new languages, Chinese and Arabic, for the CoNLL-2012 Shared Task on the Multilingual Coreference Resolution. Our team has some previous experience with extending BART to cover languages other than English, in particular, Italian and German. For those languages, however, most of our team members had at least an advanced knowledge, allowing for more straightforward engineering and error analysis. Both Arabic and Chinese present a challenge: they require new mention detection algorithms, as well as special language-dependent techniques for extracting mention properties.

For Arabic, we have proposed several simple adjustments to extract basic morphological information. As our experiments show, this can potentially lead to a substantial improvement. The progress, however, is hindered by the mention detection quality: even though our TKEMD module outperforms the lower bound baseline, there is still a lot of room for improvement, that can be achieved after a language-aware error analysis.

For Chinese, the subtask of extracting relevant linguistic information has turned out to be very challenging. We believe that, by elaborating on the methods for assigning linguistic properties to nominal mentions and combining them with the TKEMD module, one can boost the performance level of a coreference resolver.

7 Acknowledgments

The research described in this paper has been partially supported by the European Community's Seventh Framework Programme (FP7/2007-2013) under the grants #247758: EternalS – Trustworthy Eternal Systems via Evolving Software, Data and Knowledge, and #288024: LiMoSINe – Linguistically Motivated Semantic aggregation engiNes.

References

Olatz Arregi, Klara Ceberio, Arantza Díaz De Illarraza, Iakes Goenaga, Basilio Sierra, and Ana Zelaia. 2010. A first machine learning approach to pronominal anaphora resolution in Basque. In *Proceedings of the 12th Ibero-American conference on Advances in artificial intelligence*, IBERAMIA'10, pages 234–243, Berlin, Heidelberg. Springer-Verlag.

Elizabeth Baran and Nianwen Xue. 2011. Singular or plural? Exploiting parallel corpora for Chinese number prediction. In *Proceedings of the Machine Translation Summit XIII*.

Shane Bergsma and David Yarowsky. 2011. NADA: A robust system for non-referential pronoun detection. In *Proceedings of the Discourse Anaphora and Anaphor Resolution Colloquium*, Faro, Portugal, October.

Michael Collins. 1999. *Head-Driven Statistical Models for Natural Language Parsing*. Ph.D. thesis, University of Pennsylvania.

Sanda Harabagiu and Steven Maiorano. 1999. Knowledge-lean coreference resolution and its relation to textual cohesion and coherence. In *Proceedings of the ACL Workshop On The Relation Of Discourse/Dialogue Structure And Reference*.

Sanda Harabagiu, Răzvan Bunescu, and Steven Maiorano. 2001. Text and knowledge mining for coreference resolution. In *Proceedings of the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 55–62.

Frederick Hoyt. 2008. The Arabic noun phrase. In *The Encyclopedia of Arabic Language and Linguistics*. Leiden:Brill.

Zhiheng Huang, Guangping Zeng, Weiqun Xu, and Asli Celikyilmaz. 2009. Effectively exploiting WordNet in semantic class classification for coreference resolution. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Ryu Iida and Massimo Poesio. 2011. A cross-lingual ILP solution to zero anaphora resolution. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 804–813.

Alessandro Moschitti. 2006. Efficient convolution kernels for dependency and constituent syntactic trees. In *Proceedings of European Conference on Machine Learning*, pages 318–329.

Alessandro Moschitti. 2008. Kernel methods, syntax and semantics for relational text categorization. In *Proceeding of the International Conference on Information and Knowledge Management*, NY, USA.

Simone Paolo Ponzetto and Michael Strube. 2006. Exploiting semantic role labeling, WordNet and Wikipedia for coreference resolution. In *Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 192–199.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In *Proceedings of the Sixteenth Conference on Computational Natural Language Learning (CoNLL 2012)*, Jeju, Korea.

Marta Recasens, Lluís Màrquez, Emili Sapena, M.Antònia Martí, Mariona Taulé, Véronique Hoste, Massimo Poesio, and Yannick Versley. 2010. SemEval-2010 Task 1: Coreference resolution in multiple languages. In *Proceedings of the 5th International Workshop on Semantic Evaluations (SemEval-2010)*, Uppsala, Sweden.

Sriparna Saha, Asif Ekbal, Olga Uryupina, and Massimo Poesio. 2011. Single and multi-objective optimization for feature selection in anaphora resolution. In *Proceedings of the International Joint Conference on Natural Language Processing*.

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

Olga Uryupina and Massimo Poesio. 2012. Domain-specific vs. uniform modeling for coreference resolution. In *Proceedings of the Language Resources and Evaluation Conference*.

Olga Uryupina, Sriparna Saha, Asif Ekbal, and Massimo Poesio. 2011. Multi-metric optimization for coreference: The UniTN / IITP / Essex submission to the 2011 CONLL shared task. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning (CoNLL 2011)*.

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 *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies*, pages 9–12.