Domain-specific vs. Uniform Modeling for Coreference Resolution

Olga Uryupina[1], **Massimo Poesio**[1,2]

[1] University of Trento, Center for Mind/Brain Sciences,[2] University of Essex, Language and Computation Group
uryupina@gmail.com, poesio@essex.ac.uk

Abstract

Several corpora annotated for coreference have been made available in the past decade. These resources differ with respect to their size and the underlying structure: the number of domains and their similarity. Our study compares domain-specific models, learned from small heterogeneous subsets of the investigated corpora, against uniform models, that utilize all the available data. We show that for knowledge-poor baseline systems, domain-specific and uniform modeling yield same results. Systems, relying on large amounts of linguistic knowledge, however, exhibit differences in their performance: with all the designed features in use, domain-specific models suffer from over-fitting, whereas with pre-selected feature sets they tend to outperform union models.

Keywords: coreference, discourse, domain adaptation

1. Introduction

In many areas of Computational Linguistics it has been shown that domain selection and adaptation methods have the potential to improve the performance of systems. The issue is of particular interest for research in coreference as many of the corpora used for evaluation—e.g., the ACE corpora—consist of several domains, which however are not particularly big in size, so that it's an often considered question whether one would get a better performance by training and testing separately or putting all data together.

And it's becoming even more of an issue considering that the biggest resource currently available, OntoNotes-3, consists of a number of texts from different domains, so that finding homogeneous subsets may be very useful (as well as being practically necessary given the size of the corpus).

No systematic study of the effect of diversity on developing coreference methods has been carried out yet, but there has been some preliminary work investigating the effect of corpus composition on the performance of coreference algorithms, showing that even apparently similar corpora like MUC and ACE in fact differ in a number of ways (Stoyanov et al., 2009).

As a first step in this direction, we propose in this paper a measure of corpus homogeneity w.r.t. coreference that can be used to test whether it's sensible to blend together different domains at training, or it's best to keep them separate. We believe that this measure might help us better understand the structure of the coreference datasets and find homogeneous clusters within them, ultimately improving the resolution accuracy.

In this work we also remain aware that corpus homogeneity may have different effects on different types of coreference feature sets - in particular, that shallow feature sets, which have been shown to be less affected by the size of the training corpus (Soon et al., 2001; Uryupina, 2006) may also be less sensitive to corpus homogeneity.

In the present paper we investigate possibilities of improving the system's performance through learning domain-specific models. Our hypothesis is that a classifier, relying on a rich set of linguistic features, can benefit from capturing domain specific information. We run our experiments on three corpora: ACE-02, ARRAU and OntoNotes. These datasets differ with respect to their size, domain diversity and annotation guidelines and provide therefore a good testbed for our approach.

2. Datasets

2.1. Corpora

In the present study, we compare three datasets: ACE-02 (Doddington et al., 2004), ARRAU (Poesio and Artstein, 2008) and OntoNotes (Pradhan et al., 2011). All these corpora contain documents from different domains, and the corresponding information has been preserved in the distributions. All the ACE-02 and OntoNotes documents are news, whereas ARRAU contains news, medical texts, fiction and dialogues.

In Table 1 we provide some statistics for the training data in the whole datasets and their domains: the average length of a document, total number of tokens and mentions. It is clear that "domains" in the three corpora are composed very differently: thus, the nw OntoNotes domain is larger than the whole ACE-02 set.

The corpora have been annotated according to very different guidelines. The ACE-02 dataset has been developed mainly from the Information Retrieval perspective, whereas both the ARRAU and OntoNotes annotation schemes represent a more linguistic view of coreference.

The ACE guidelines focus only on specific predefined entity types: PERSON, ORGANIZATION and so on. The OntoNotes data, on the contrary, contain annotations of all the entities. ACE mentions contain information on their *minimal spans*, that allow for less restrictive alignment of system vs. gold boundaries. OntoNotes mentions are expected to be recovered in their exact boundaries. The ARRAU guidelines require all the mention to be annotated, but the minimal spans are provided only for some domains (rst, vpc).

The ACE guidelines follow a very rough definition of coreference, introduced at the MUC initiative, that has since been criticized a lot by linguists (van Deemter and Kibble, 2001). The OntoNotes scheme follows a more subtle line, introducing a distinction between "Identical" and "Appositive" coreference. In ARRAU, cases of "appositive" coreference are labeled as "non-referring". The guidelines

	avg doc length	tokens	mentions	C$_{dom}$
ARRAU				
gnome	4083	12250	3550	0.14
pear	751	11270	3126	0.11
rst	733	108609	33843	0.25
t91	892	10704	2132	0.16
t93	637	10198	2085	0.16
vpc	1062	21248	6592	0.23
whole	792	155976	45585	0.31
ACE-02				
bnews	314	67720	10086	0.52
npaper	950	71230	11320	0.38
nwire	629	81767	10868	0.42
whole	525	221138	32274	0.47
OntoNotes/CoNLL				
bc	517	142692	37267	0.19
bn	239	182270	51358	0.22
mz	403	164632	43792	0.16
nw	521	387827	103152	0.19
wb	755	131338	36039	0.15
whole	426	1008759	271608	0.21

Table 1: Train data statistics for 3 corpora: ARRAU, ACE-02 and OntoNotes.

also take different views on pronouns (especially generic “you”), generic nouns (including bare plurals), coordinations and pre-modifiers. Singletons (referring noun phrases that do not participate in any coreference relations) are annotated in ACE-02 and ARRAU, but not in OntoNotes.

All these differences make it infeasible to train a model on one corpus and then test it on another one: the discrepancies in the annotation guidelines would make it a futile exercise. We therefore do not follow the common practice in domain adaptation studies, where corpora are created by merging several resources (Daume, 2007; McClosky et al., 2010; Plank and van Noord, 2011), but focus on sub-domains of the investigated corpora.

2.2. Measuring domain homogeneity

Both ACE-02 and OntoNotes contain news documents. This raises a question of the applicability of the notion of “domain” to these datasets. To identify documents with similar discourse properties with respect to coreference, we compare distributions of their nominal mentions (i.e. basic units for any coreference resolution system) across different categories. For example, a long document, containing a lot of pronouns, is not similar to a short snippet full of proper names. A coreference resolution system should rely more on salience for the former and on matching for the latter. It might therefore be beneficial to train a coreference classifier on domain-specific documents, exhibiting similar discourse properties.

We have investigated a number of indicators to quantify the document structure, partially motivated by Stoyanov et al. (2009). Each indicator is a ratio of the number of some specific mentions in the given document (for example, pronouns) to the total number of mentions. By definition, all the indicator take values within the $[0..1]$ range. Table 2 shows the indicators used in the present study. We determine whether a mention is a pronoun, a name or a nominal automatically, using simple heuristics (indicators $2, 3, 4$). For indicators $5, 6, 7$, we extract semantic class labels for our mentions from either CARAFE (ACE02) or the Stanford NER toolkit (ARRAU, OntoNotes). Finally, indicators $8, 9$ and 10 use information on the structure of coreference chains in a document (for example, our indicator 10 measures the number of singleton chains). In the present study, we only use our indicators for the training data, so we rely on the gold annotation here.[1] In our future work we aim at more fine-grained analysis.

1. # mentions per token
2. # pronouns per mention
3. # names per mention
4. # nominals per mention
5. # PERSONs per mention
6. # ORGANIZATIONs per mention
7. # LOCATIONs+# GPEs+# GSPs per mention
8. # coreference-chains per mention
9. size-of-longest-chain / # mentions
10. # singletons per mention

Table 2: Indicators of document structure w.r.t. coreference

For each domain, we extract the indicators for each document and then compute the centroid. The average distance to the centroid across the documents is then used as a measure of domain homogeneity C_{dom}, also reported in Table 1.

Values of C_{dom} depend on the annotation scheme. For example, the OntoNotes guidelines assume no singletons and thus one of the indicators is always 0 and does not contribute to C_{dom}. It is therefore impossible to compare C_{dom} values across corpora. However, we can use C_{dom} to investigate domains within the same corpus.

Table 3 shows distances between centroids for different domains of our three corpora. For ACE, differences within domains (C_{dom}, last column of Table 1) are larger than those between domains (for example, the distance between the centroids for $npaper$ and $nwire$ is 0.16, much smaller than the average for both domains). This is also reflected in the relatively low C_{dom} value for the whole ACE-02 corpus. In OntoNotes, some domains (bc) are clearly defined, whereas others are very similar (the distance between mz and wb is 0.09). The C_{dom} value for the whole corpus is just slightly above those for its individual domains. Finally, the domain structure of ARRAU is successfully captured by our indicators: we see a cluster for dialogues ($t91$ and $t93$), a cluster for news (rst, vpc), medicine ($gnome$) and fiction ($pear$). All those clusters are distinct: for example, the distance between $pear$ and rst is 0.65. The C_{dom} value for the whole corpus confirms that ARRAU is a very heterogeneous dataset, comprising distinct domains.

[1] In the future work, we plan to apply our indicators to classify unseen documents into appropriate domains — in this case we will have to exclude indicators $8, 9$ and 10.

ACE-02

	npaper	nwire
bnews	0.24	0.16
npaper		0.17

OntoNotes/CoNLL

	bn	mz	nw	wb
bc	0.21	0.23	0.29	0.17
bn		0.16	0.15	0.21
mz			0.11	0.09
nw				0.19

ARRAU

	pear	rst	t91	t93	vpc
gnome	0.47	0.26	0.42	0.42	0.17
pear		0.65	0.63	0.65	0.58
rst			0.44	0.44	0.10
t91				0.05	0.44
t93					0.44

Table 3: Distances between domains for 3 corpora: ACE-02, OntoNotes and ARRAU

3. Methodology

For our experiments, we use BART (Versley et al., 2008), a toolkit for coreference resolution. At the recent CoNLL shared task, it has shown reliable performance, with two independent BART-based submissions ranking both among the top systems.

We view coreference resolution as a binary classification problem (Soon et al., 2001). Each classification instance consists of two mentions, i.e. an anaphor and its potential antecedent. Instances are modeled as feature vectors and are handed over to a binary classifier that decides, given the features, whether the anaphor and the candidate antecedent are coreferent or not. All the feature values are computed fully automatically, without any manual intervention. We learn with the J48 Decision Trees classifier provided with the Weka package. It has been chosen for technical reasons: given the size of the OntoNotes dataset, it is virtually impossible to train any other learner. We have also rerun our experiments on ACE-02 with the Maximum Entropy and have observed very similar performance (both in trends and absolute values).

All the experimental results are reported on system mentions. We extract ACE-02 mentions using CARAFE[2]. For ARRAU and OntoNotes, we have developed a complex heuristic for extracting mentions from parse trees: first merging NP constituents and named entities and then filtering the resulting list to discard expletives and similar expressions (cf. Uryupina et al. (2011) for details).

We compare the system performance for different feature sets. First, we evaluate a knowledge-poor approach, that relies on a dozen of features advocated by Soon et al. (2001). Second, we train a classifier with a linguistically motivated set of 42 features (cf. Uryupina et al. (2011) for details). We have observed, however, that this classifier performs moderately on smaller training sets. We attribute this behavior to the over-fitting problem. To alleviate it, we have applied a very basic form of feature selection, reducing the feature set to its half. For ACE-02, we use the feature set optimized on the $nwire$ domain. For ARRAU and OntoNotes, as they cover similar phenomena, we use the feature set optimized on OntoNotes. This is a very naive solution that only aims to reduce the amount of over-fitting. We will attempt a proper domain-specific feature selection in our future experiments.

4. Results and Discussion

Table 4 shows our experimental results (MUC F-score). For each feature set, we compare models trained on specific domains against those trained on the whole corpus.

For knowledge-poor models, the investigated strategies yield very similar scores with no significant differences between the conditions. This can be explained by two factors. First, knowledge-poor models have no access to any domain-specific information: with very basic surface-level features, the system cannot capture fine linguistic phenomena. Training on a specific domain, therefore, does not give any advantage. Second, such models do not require much training data: for example, Soon et al. (2001) report that their system achieved nearly its expected performance already when trained on 5 documents. Training such a model on the whole corpus doesn't bring any advantage either.

A rich linguistically-motivated model, relying on a large feature set, could be expected to capture more complex properties and thus benefit from domain-specific resolution. The experimental results, however, show that such models suffer from over-fitting: in both conditions, the performance goes down compared to the simple knowledge-poor model, especially for ARRAU and OntoNotes.[3] When we train such a model on a smaller domain-specific dataset, the issue becomes more crucial, resulting in a further drop in performance. This can be clearly seen with the ACE domains. The OntoNotes domains are larger and therefore the over-fitting problem is less pronounced. The only exception to this trend are some of the ARRAU domains that are very distinct from the rest of the corpus.

At the first glance, these results suggest that knowledge-rich models are not robust enough and should be replaced with surface-level algorithms, similar to Soon et al. (2001). It has been shown, however, that linguistically-motivated approaches can benefit a lot from feature selection (Ng and Cardie, 2002; Hoste, 2005; Uryupina et al., 2011): with a smaller set of carefully chosen features, we can expect to capture the most important properties of coreference without too much over-fitting. The rightmost columns of Table 4 compare the performance level of such a model for the same experimental conditions. For both ACE and OntoNotes, such approach yields the best results, outperforming both the surface-level and the all-features systems. For ARRAU, however, the results are still moderate. This highlights the importance of proper corpus- and domain-specific feature selection. With appropriate smaller feature

[2] `http://sourceforge.net/projects/carafe`

[3] Our features have been originally designed for the MUC and ACE corpora and cannot be expected to achieve their optimal performance on other datasets.

	Soon et al		all features		Selected features	
	domains	union	domains	union	domains	union
ARRAU						
gnome	58.06	56.92	56.12	51.56	56.38	56.11
pear	66.74	67.36	64.35	65.06	66.29	65.24
rst	59.51	59.36	53.61	53.77	56.88	57.97
t91	39.01	36.85	39.59	35.83	37.86	36.67
t93	43.17	42.9	41.07	39.14	47.55	43.31
vpc	57.3	57.58	52.36	54.54**	54.91	57.73
whole corpus	56.66	56.04	52.92	51.91	54.84	55.29
ACE-02						
bnews	69.33	71.33	74.15	75.46	73.17	72.61
npaper	70.92	69.69	70.93	72.37**	72.53**	67.61
nwire	69.43	69.08	67.85	72.03**	76.08**	72.96
whole corpus	69.91	70.05	70.9	73.29**	73.97**	71.12
OntoNotes/CoNLL						
bc	55.04	55.62	54.76**	53.85	60.71	59.52
bn	56.42	56.33	53.71	52.93	58.39**	58.17
mz	59.56	60.2	52.93	53.35	61.65	62.42
nw	52.03**	51.53	44.24	44.63**	55.5	54.47
wb	51.07	53.05	47.39	46.61	53.91	53.36
whole corpus	54.17	54.5	49.55	49.35	57.74**	57.05

Table 4: Experimental results (MUC F-score) for 3 corpora: ARRAU, ACE-02 and OntoNotes. Significant differences between domain and union modeling shown with ** (sign test on per-document scores, $p < 0.05$). NB: sign test not applicable to the ARRAU $t91$, $t93$, $gnome$ and $pear$ domains due to the insufficient test set size.

sets, the domain-specific modeling brings a significant advantage over the union one.[4]

5. Conclusion

In this paper we have investigated possibilities of domain-specific modeling for coreference resolution. We have shown that for modern knowledge-rich algorithms it might be beneficial to split the training set into different domains to be processed separately, provided the system relies on efficient feature selection techniques to overcome the overfitting problem. This holds even for corpora containing relatively similar documents (ACE-02, OntoNotes). It must be noted that a domain-specific approach is only beneficial when the corresponding domain contains a sufficient number of documents. For very small domains, for example ARRAU vpc, one should consider merging the training set with other similar domains (using the distance between the centroids of indicator vectors as a similarity measure) or simply retract to the union modeling.

In our study we rely on a very naive measure of document similarity with respect to coreference. In the future work, we plan to elaborate on this measure, incorporating more indicators of the discourse structure and combining it with other metrics of document similarity (e.g. topic models).

Our analysis shows that some documents might belong to one domain but be more similar to another one (e.g., the ACE-02 domains provide only very poor clusters). This suggests that, for a given test document, it might be beneficial to construct its domain automatically, ignoring the split suggested in the corpus distribution (cf. Plank and van Noord (2011) for a similar approach to parsing). We plan to pursue this line in our future work.

6. Acknowledgements

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.

7. References

Hal III Daume. 2007. Frustratingly easy domain adaptation. In *Proc. ACL 2007*.

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassell, and Ralph Weischedel. 2004. The automatic content extraction (ACE) program–tasks, data, and evaluation. In *Proc. LREC-04*.

Veronique Hoste. 2005. *Optimization Issues in Machine Learning of Coreference Resolution*. Ph.D. thesis, Antwerp University.

David McClosky, Eugene Charniak, and Mark Johnson. 2010. Automatic domain adaptation for parsing. In *Proc. NAACL 2010*.

Vincent Ng and Claire Cardie. 2002. Improving machine learning approaches to coreference resolution. In *Proc. ACL-02*, pages 104–111.

[4] An exception is a small ARRAU vpc domain: as our indicators show, it is very similar to a much larger domain, rst, which makes it beneficial to train a linguistically-rich model on the whole corpus.

Barbara Plank and Gertjan van Noord. 2011. Effective measures of domain similarity for parsing. In *Proc. ACL 2011*.

Massimo Poesio and Ron Artstein. 2008. Anaphoric annotation in the arrau corpus. In *Proc. LREC-08*.

Sameer Pradhan, Lance Ramshaw, Mitchell Marcus, Martha Palmer, Ralph Weischedel, and Nianwen Xue. 2011. Conll-2011 shared task: Modeling unrestricted coreference in ontonotes. In *Proc. CoNLL 2011*, Portland, Oregon, June.

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.

Veselin Stoyanov, Nathan Gilbert, Claire Cardie, and Ellen Riloff. 2009. Conundrums in noun phrase coreference resolution: making sense of the state-of-the-art. In *Proc. ACL 2009*.

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 *Proc. CoNLL-2011*.

Olga Uryupina. 2006. Coreference resolution with and without linguistic knowledge. In *Proc. LREC-06*.

Kees van Deemter and Rodger Kibble. 2001. On coreferring: Coreference in muc and related annotation schemes. *Computational Linguistics*.

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 *Proc. ACL-08*, pages 9–12.