Exploring Co-Reference Chains for Concept Annotation of Domain Texts

Petya Osenova, Laska Laskova, Kiril Simov
Linguistic Modelling Department, IPP, Bulgarian Academy of Sciences
Acad. G.Bonchev 25A, 1113 Sofia, Bulgaria
E-mail: petya@bultreebank.org, laska@bultreebank.org, kivs@bultreebank.org

Abstract

The paper explores the co-reference chains as a way for improving the density of concept annotation over domain texts. The idea extends authors' previous work on relating the ontology to the text terms in two domains – IT and textile. Here IT domain is used. The challenge is to enhance relations among concepts instead of text entities, the latter pursued in most works. Our ultimate goal is to exploit these additional chains for concept disambiguation as well as sparseness resolution at concept level. First, a gold standard was prepared with manually connected links among concepts, anaphoric pronouns and contextual equivalents. This step was necessary not only for test purposes, but also for better orientation in the co-referent types and distribution. Then, two automatic systems were tested on the gold standard. Note that these systems were not designed specially for concept chaining. The conclusion is that the state-of-the-art co-reference resolution systems might address the concept sparseness problem, but not so much the concept disambiguation task. For the latter, word-sense disambiguation systems have to be integrated.

1. Introduction

Domain texts, annotated with the key conceptual information in the chosen domain, are a necessity for applications, such as information retrieval, information extraction, life-long learning, question answering, etc.

In our previous work, we relied on an ontology-to-text relation model in the annotation process. It provides a mechanism for explicating the conceptual information within the text. The current ontology-to-text relation model comprises a domain ontology; a lexicon, mapped to it; and a concept annotation grammar based on cascaded regular grammar technology, which finds the concept lexicalizations in the text, and assigns to them the appropriate concepts from the ontology – for more details, see (Osenova, Simov, and Mossel 2008) and (Simov and Osenova 2008).

However, the current implementation detected a problem, which is the concept sparseness of the annotation (about 2 domain concepts per sentence). This is far from enough for observing a conceptual network over a text, and for evaluating the concept distribution. For that reason, we decided to enhance the implicit domain semantic information through co-reference relations. Co-reference chains are targeted as additional context pointers for a concept within the concept annotated domain texts. Thus, the co-reference resolution systems and the concept automatic annotator need to reach a common ground and to start working together for better conceptual coverage over the texts.

For this purpose, we performed the following test case workflow: first, manual annotation with co-references of texts in IT domain as a gold standard, and then, testing of two automatic systems for co-reference annotation over the same texts.

The challenge here is the attempt in establishing relations between the co-reference mechanisms and the ontological concepts with the idea to pass the conceptual information from an annotated concept lexicalization to its co-reference expressions in the text. This is in contrast to most popular works in NLP, which focused on chaining the named entities, synonymy and anaphora. In this sense, our task is not trivial.

Pursuing the relation between concept annotation and co-references in general is not new. It has been approached from various perspectives, but with the aim to improve the co-references. For example, (Lech and de Smedt 2006) and (Nikolov et. al 2009), among others, exploit the semantic features from ontology in order to improve the co-reference chaining; (Kawazoe et al. 2003) designed a software that helps experts in biomedical domain to create ontologies and annotate texts with co-references. In our case study, we adopted the ideas in these papers (together with the work on anaphora and co-reference annotation in general). In the future work, we intend to apply these combined approaches for the implementation of a new version of the ontology-to-text relation model.

The paper is structured as follows: section 2 describes the co-reference mechanisms with respect to the concept annotated texts. Section 3 highlights the characteristics of the corpus. Section 4 describes the manual annotation layer. Section 5 reports the experiments with two state-of-the-art automatic systems. Section 6 outlines the case study evaluation and results. Section 7 concludes the paper.

2. The concept annotation process and co-references

The concept annotation is based on a domain ontology in IT domain. As mentioned above, it relies on a model that connects the ontology, the lexicon and the text.

Semantic retrieval depends very much on both measures – recall and precision of the annotation. Our previous work showed that the concept annotation based mainly on the terms from the lexicon is rather sparse. The paradigmatic relations, such as *is-a, part-of, used-for, composed-of*, can be detected through the ontological hierarchy. However,

the syntagmatic ones, such as lexical chains and anaphors, remain implicit. Under a lexical chain we mean the usage of a more general term as a substitute to the specific one (*page* for *web page*). Under anaphor the standard notion is meant - using an anaphoric pronoun as a reference to the concept (*web page - it*).

The members of a lexical chain would receive different concept labels from the ontology. For the above example pair, the term *web page* would receive the subconcept label, while the term *page* – the more general one. However, in the context, they are realizations of the same concept.

The pronoun members of anaphoric relations are not considered at all by the ontology, being just referring words without its own content.

The ambiguity is caused mainly by the general concepts. For example, the concept *page* might go to *text page* and to *web page*. If the co-reference linkage is active, then disambiguation problem might also be resolved.

To sum up, the semantic retrieval loses from the unresolved concept ambiguity and the missing connections within the context.

Pure repetitions might be a challenge if parts of various concepts. But this issue is a more ontology coverage problem than a text occurrence issue.

For that reason, we have decided to explore the potential of the co-references, and more precisely, the co-reference systems, in a case study for two purposes: disambiguation of the ambiguous concepts, and providing more syntagmatic contexts for the concepts in the retrieval results.

3. Corpus sample as a gold standard

The complete manually annotated corpus is in English and it comprises documents on two specific mark-up languages – XML and HTML. It contains 158 769 tokens and 24 688 domain specific concepts, of which 4149 participate in a concept chain. The co-reference annotation was performed on the top of the concept annotation. In the table below the percentage of the concept-receivers as well as the percentage of the concept specialized meaning within the concept chains is presented:

	concept-receivers	concept-specialization
%	57.82%	45.90%

As it can be seen, the role of the co-reference chaining for concept transfer in a domain text is substantial.

The share of new concept elements, becoming explicit from the annotation process, is 31.33 % (1300).

However, for the experiment with the automatic systems, a single document on HTML was chosen, which comprises 10 205 tokens. From all the tokens in this document, 6350 met the preliminary condition to become a markable candidate, that is, they are not function words, punctuation marks, interrogative pronouns or verb forms. Only 1330 of them turned out to be concept bearers. Altogether, there are 92 concept chains covering the content of 25 concepts.

Since we were interested only in chains, which included lexicalizations for the concepts from our IT ontology, not all existent in the text concept chains have been marked. Thus, 273 expressions were co-indexed: 33.70% concept bearers (antecedents), 24.90% pronouns, 41.39% content words that receive a (new) concept as a result of being an element of a chain.

4. Annotation Strategy

According to our model, each detected lexicalization receives the equivalent concept label as well as its super-concept label in the background. For example, *HTML editor* has a super-concept *Word Processor* or *HTML tag* has a super-concept *Tag*. While in the first case it is more unlikely to use the general concept instead of the more specific one in the text, in the second case this is very likely. For that reason, when establishing the lexical or anaphoric chain, all concept-bearers and concept-receivers share the same index. Concept-bearers are the terms that receive their labels from the ontology. Concept-receivers are those expressions, which get the label from a concept-bearer, based on the lexical chain or anaphoric relation. Thus, when participating in a chain, a text item can have or might not have a concept label from the ontology, but it obligatorily has a context-bound concept label, being part of the chain. By default we use only the *equivalence* relation, which corresponds to the relation IDENT(ity) in the MUC annotation schema (Chinchor 1998). In very rare cases, a chain among concept-superconcept are also considered (see Section 4). The MUC SGML structure of identity is given below for clarity and comparison with our representation:

```
<COREF @ID="unique_number_for_antecedent">
                antecedent_phrase
</COREF>
<COREF @ID="unique_number_for_anaphora"
      @TYPE="IDENT"
      @REF="unique_number_for_antecedent">
                anaphora_phrase
</COREF>
```

Since our annotation scheme is designed in XML, it adheres to the following rules: The elements in the chain are marked as the element - <Concept>. They receive the same attribute - @index. The concept bearer's attribute - @class - is predefined. Anaphoric concepts, which are concept-receivers, may or may not have this @class attribute (i.e. concept annotated on the basis of the lexicon), but all of them receive the context attribute - @c-class (i.e. concept as bounded by the context in the chain). Its value is determined by the anaphoric relation with the antecedent. In accordance with the co-reference ideology, we considered NPs as possible markable candidates, including phrases with elliptical heads, relative pronouns, personal and possessive, reflexive and demonstrative pronouns. The XML structure is as follows:

```
<Concept @index="in#id"
          @class="ontology_concept_original"
          @c-class="ontology_concept_received">
      <tok1>lexical_term</tok1>  …….
      <tokn>lexical_term</tokn>
</Concept>
```

The boundaries of a concept chain are fixed by the next appearance of an antecedent, which bears the same concept in the discourse.
Here come several scenarios concerning the relation between concept-bearers and concept-receivers.
If a potential concept-bearer does not have the attribute @class (i.e. the concept is not domain specific or not present in the ontology), the further candidates for concept-receivers are ignored, and chains are not formed. At the same time, the term is registered for being added to the ontology. In the process of work, it turned out that the missing concepts are very specific. They represent a subdomain in the domain. For the whole corpus around 150 new candidate concepts have been detected and added to the ontology. For example, the term *HTML address tag* is a subtype of *HTML tag*, present in the ontology.
Not all concepts with the same concept label from the ontology are co-indexed to participate in the same chain. This happens when the term is assigned the more general concept (*page*), but it actually refers to the more specific one (*web page*). In the same text, the term *page* might be used in both senses.
Another scenario is the anaphoric chain occurrence, which happens to be a frequent phenomenon in the text cohesion. Needless to say, when there is an anaphoric relation between a pronominal expression and a concept-bearer, the anaphora is also annotated with a concept, whose value is identical to the antecedent. Thus, both of them receive @index attribute with equal value. Let us consider the following anaphoric sequence, where *XML* transfers its concept label to the pronoun *it*:
***XML** is used to aid the exchange of data. **It** makes it possible to define data in a clear way.*
The structure is as follows:

<Concept @class="http://www.lt4el.eu/CSnCS#XML"
@index="in001">
<tok>***XML***</tok>
</Concept>
is used to aid the exchange of data.

<Concept @c-class="http://www.lt4el.eu/CSnCS#XML"
@index="in001">
<tok>***It***</tok>
</Concept>
makes it possible to define data in a clear way.

The context-dependent attribute (@c-class) for the pronominal expression indicates the transferred and already common-shared concept (underlined).
In the non-co-reference concept annotation, the added value of the information about XML, presented in the anaphoric chain, would have been lost. But in this scenario it contributes to the concept description.
In case of concept disambiguation, the annotation procedure is the same, except for the fact that the anaphoric expression (in this case - lexical NP) has both attributes – the label, assigned from the ontology (@class) and the one, received within the chain (@c-class). The example below is again with the token *page*.

Let us consider the sentence:
***HTML file** can link to an external style sheet and also include a style element for additional style settings specific to this **page**.*
Here the second occurrence of *page* is bound co-referentially by the concept-bearer *HTML file* and its ontological label *HTML Page*.

<Concept
@class="http://www.lt4el.eu/CSnCS#HTMLPage"
@index="in007">
<tok>***HTML***</tok>
<tok>***file***</tok>
</Concept>

can link to an external style sheet and also include a style element for additional style settings specific to this

<Concept
@class="http://www.lt4el.eu/CSnCS#Page"
@c-class="http://www.lt4el.eu/CSnCS#HTMLPage"
@index="in007">
<tok>***page***</tok>
</Concept>

Since the problems with the annotation include also partial concept detections, or more precisely, concepts that are part of other concepts, they are places for artificial ambiguities along with the genuine ones (among domain and non-domain terms; among various domain senses of a domain term). Thus *page* might happen also to be wrongly recognized as the general concept *page* inside a more specific term (*web page*), if this term was not mapped into a concept in the ontology.
There is a very limited number of cases where the value for the anaphoric context class is actually a super concept for the antecedent's concept-bearer. In such cases no transferring is performed, only chaining. For example, in the sentence: ***Ordered lists** are **ones**, where the browser numbers each successive list item starting with '1'*, the anaphoric 'ones' refer to the more general term:

<Concept
@class="http://www.lt4el.eu/CSnCS#NumberedList"
@index="in002">
<tok>***Ordered lists***</tok>
</Concept>
are
<Concept
@c-class=" http://www.lt4el.eu/CSnCS#List'
@index="in002">
<tok>***ones***</tok>
</Concept>
where the browser numbers each successive list item starting with "1."

Following these principles, we have annotated a domain corpus of more than 150 000 tokens for future observations and tests.

5. The Automatic settings

Our first attempt to address the co-referential task in a concept-based framework was to exploit off-the-shelf systems as they are distributed by their developers. Experiments have been made with several other systems, but here we report only the results from both most successful for our purposes ones – OpenNLP and BART.

Since none of the tools was designed specially for concept chaining, but they rather handled various types of co-reference resolution, it would be unfair to evaluate their results directly against the golden standard annotation. Therefore, the common measures like precision, recall and F-measure were not used. Instead, the systems were evaluated against the fact to what extent they could improve the concept coverage via concept transfer within the lexical or anaphoric chains.

OpenNLP [1] is a well-known Java-based toolkit that performs all standard NLP steps (sentence splitting, tokenization, POS-tagging, etc.), including co-reference detection, that makes use of WordNet.

BART [2] (Beautiful/Baltimore Anaphora Resolution Toolkit) is an open source modular toolkit developed as a result of the project Exploiting Lexical and Encyclopedic Resources For Entity Disambiguation 2007. It includes ideas from GuiTAR system and other co-reference systems. BART architecture allows for further exploration of different pre-processing and resolving methods. Both input and output are in XML format (MMAX2 format). BART can be used as a platform for experimentation or as a off-the-shelf tool for anaphora resolution. On MUC-6 corpus BART had better performance in pronoun resolution than JavaRAP (Versley et. al. 2008).

However, these two systems (as all other ones in NLP world) have been tuned to specific domains and/or tasks. Thus, their adoption was not straightforward and easy for the IT domain – just to mention some stumbling-stones: visual means of content structuring that could not be taken into account when building the discourse structure solely on textual indicators, incorporation of pieces of HTML, XML or Java code within the texts, ambiguity of highly specific terms, common lack of token and type distinction.

6. Results and evaluation

The two systems were run on the chosen HTML file with their default settings. This is our baseline for the further experiments. First, they detected the markables, and then – performed the chainings. The concept annotation was hidden to them. It was used only in the evaluation from the automatic co-reference resolution.

The number of co-reference chains, marked by OpenNLP, is 154. Compared to the manually tagged elements, OpenNLP markables are often maximal NPs, which is in agreement with the MUC annotation scheme requirements. Approximately one quarter of them (24.67%) are expressions (usually heads in an NP) related to a concept from the domain ontology. Only 1 of the chains could be used for sense disambiguation (*web page – my page*); 50% have as their members pronouns, and the rest are lexical repetitions. Thus, the performance of OpenNLP is very close to the manual work observations. The extension of the concepts in the text due to anaphoric chaining is 50 %, which is a promising start.

Based on these results, we can draw the conclusion that OpenNLP might be used as a means to detect the context-dependent meaning of the pronouns, which denote domain specific concepts. This in turn would provide a more adequate picture of the text saliency for the different concepts in the analyzed document.

The output from BART includes 373 co-reference chains and compared to the OpenNLP output, there are more cases of embedded markables, e.g. {2the {1browser} window}, where the term *browser* is embedded in another term - *browser window*. Taking into account the results from the previous experiment with OpenNLP, we expected that the co-reference information provided by BART might also better support anaphora resolution type of concept chaining that lexical one (excluding the pure repetitions). This assumption was confirmed. However, most of the chains include repetitions of one or two expressions. For example, one of the chains contains 131 markables, 28 of them personal pronouns ("it"), 2 possessive ("its") and the rest are abbreviation tokens ("HTML") or chunks, including the abbreviation. Although the recall of BART is better than OpenNLP, the precision is not very good (in this example, only 2 of the pronouns were co-referential with HTML).

In the previous sections we pointed out that both recall and precision are important for the semantic retrieval. Needless to say, depending on the specific task, the former or the latter metric might become more important than the other. For the moment OpenNLP showed better results on successful expansion of concepts in the text. Thus, we included it as part of our linguistic processing pipe. Another reason is that it has potential for a fairly straightforward integration of a word sense disambiguation model.

However, since BART provides a better recall and a lot of information, other, more sophisticated settings and adaptations should be explored for our task.

7. Conclusion

Both systems, considered in our experiment setting, do not tend to take decision when there are ambiguities. In contrast to OpenNLP, BART connects named entities. However, the change of domain makes this facility an obstacle. Both systems connect only close synonyms, indicating the same concept. However, the interference of more co-reference chains fails them. Also, the systems do not connect concept–subconcept relations. BART connects all pronouns in the text, which however leads also to many undesired mistakes.

Both systems can be used for anaphora resolution, but not for disambiguation between different senses of the domain terms. For that reason, our future work on disambiguation will aim at combining co-reference

[1] http://opennlp.sourceforge.net/about.html

[2] http://www.sfs.uni-tuebingen.de/~versley/BART/

systems with word sense disambiguation ones.

For the purposes of disambiguation and better concept salience in the texts, our plans include an extension of the corpus annotation (automatically) with concepts from the top part of the ontology (in our case the Dolce – (Masolo et. al. 2002)). Thus, the non-domain lexemes would be covered, too. Then we will use this additional annotation to train the available systems for the task.

8. Acknowledgements

The work reported here is done within the context of the EU project – Language Technology for Lifelong Learning (LTfLL). We would also like to thank the three anonymous reviewers for their valuable remarks as specialists and readers.

9. References

Chinchor, N. and L. Hirschman. (1998). *MUC-7 coreference task definition, verison 3.0.* Available at: http://www-nlpir.nist.gov/related_projects/muc/proceedings/co_task.html.

Kawazoe A., T. Mullen, and K. Takeuchi. (2003). *Open Ontology Forge: A Tool for Ontology Creation and Text Annotation Applied to the Biomedical Domain*. In: Genome Informatics 14: 677-678 (2003).

Lech T. Chr. and K. de Smedt. (2006). *Enhancing Semantic Annotation through Coreference Chaining: An Ontology-based Approach*. In: Siegfried Handschuh, Thierry Declerck, Marja-Riitta Koivunen (eds.), CEUR Workshop Proceedings, Vol. 185, 2006.

Masolo, C., Borgo, S., Gangemi, A., Guarino, N., and Oltramari, A. (2002). *Ontology Library (final)*. WonderWeb Deliverable D18, December 2003. http://www.loa-cnr.it/Publications.html.

Nikolov a., V. Uren, E. Motta and A. de Roeck. (2009). *Towards instance coreference resolution in a multi-ontology environment*. Presented at: Workshop on matching and meaning, Edinburgh, UK, April 2009.

Osenova P., K. Simov, E. Mossel. (2008). *Language Resources for Semantic Document Annotation and Crosslingual Retrieval*. In: Proc. of LREC 2008, ELRA.

Simov K. and P. Osenova (2008). *Language Resources and Tools for Ontology-Based Semantic Annotation*. OntoLex 2008 Workshop at LREC 2008, pp. 9-13.

Versley, Y., Ponzetto, S.P., Poesio, M., Eidelman, V., Jern, A., Smith, J., Yang, X., Moschitti, A. (2008). *BART: A Modular Toolkit for Coreference Resolution*. ACL 2008 System demo. Available at: http://www.versley.de/.