STUDIA UNIV. BABEŞ–BOLYAI, INFORMATICA, Volume **LVI**, Number 2, 2011

ISSUES IN TOPIC TRACKING IN WIKIPEDIA ARTICLES

NATALIA KONSTANTINOVA AND CONSTANTIN ORĂSAN

1. Introduction

In the last few years, Wikipedia has become a very useful resource for NLP offering access to both structured and unstructured information that can be used for further language processing. One particularity of the Wikipedia articles is that they focus on only one topic (e.g. a product, person, location or event), which is detailed throughout the article. In order to extract comprehensive information from these articles, it is necessary to be able to track different expressions that refer to the topic. This paper discusses the issues to be tackled when a topic tracking algorithm is implemented. In order to address this problem, a shallow rule-based coreference resolution method for topic tracking was implemented.

The results of this research are intended to be used for the development of an interactive question answering (IQA) system that guides users in their search process. The answers to be provided by the IQA system will be acquired using information extraction from Wikipedia pages. To make this process more precise, it is necessary to track all the mentions of the topic throughout the article regardless of how the topic is expressed.

Attempts to use state-of-the art systems for coreference resolution showed that they provide very low precision for the task in question and link NPs which are not coreferential at all. In most cases it happens because the algorithms rely heavily on substring matching and distinguish rather poorly between entities with similar names. It can be seen very well when examining the chain generated by RECONCILE [6] for the article describing mobile phone "HTC Magic": 'The HTC Magic' - 'HTC' - 'The HTC Dream' - 'Vodafone' - 'it' - 'the Vodafone Magic'. The low performance of the state-of-the-art systems provided us with a motivation for developing our own system that will work with high accuracy for our domain.

Received by the editors: April 10, 2011.

2010 *Mathematics Subject Classification.* 68T50.

1998 *CR Categories and Descriptors.* I.2.7 [**Artificial Intelligence**]: Natural Language Processing – *Discourse*.

Key words and phrases. coreference resolution, Wikipedia, near identity.

This paper presents the first step of the research: analysis of how the topic is referred to in Wikipedia articles and which issues need to be addressed when developing a topic tracking method. Linguistic investigation of the referential expressions denoting the topic revealed that the notion of coreference is not broad enough. This issue is discussed in Section 2 with emphasis on the particularities of the Wikipedia pages. The experiment and design of evaluation are described in Section 3. The paper finishes by discussing the results of the research and conclusions.

2. The notions of coreference and near identity

Noun phrase (NP) coreference resolution is usually defined as "the task of determining which NPs in a text or dialogue refer to the same real-world entity"[3]. Coreference resolution overlaps with the field of anaphora resolution, but there is a main difference between them: anaphora is "pointing back to a previously mentioned item in the text" and coreference is "the act of referring to the same referent in the real world" [2].

The classical definition of coreference presupposes that entities can be either coreferential or not. However recent research [5] shows that this definition covers only a specific type of relation and a more fine grained definition should be used instead. We encountered the same problem while investigating a corpus of Wikipedia pages with the purpose of annotating coreference relations. One feature of Wikipedia articles is that they have a unique topic throughout the whole article, e.g. the article about "BMW E46" should focus on this model of the car. However, corpus investigation showed that it is not easy to track this topic by simply relying on the identity relation.

2.1. **Corpus annotation.**

To address the above problem, we built a corpus by extracting Wikipedia articles from the domain of products and more specifically about mobile phones and annotated them with 4 relations described below. Currently our corpus consists of 20 documents with almost 22,000 words. To enable the annotation process clear guidelines were developed to maximize the interannotator agreement. Since traditional guidelines do not cover all the situations we encountered in our domain, we had to adapt the existing guidelines [1] and change the notion of coreference. The reminder of this section briefly presents the annotation guidelines used to mark the relations of interest.

As proposed in a [1] the first step of the annotation process was to mark all the NPs, including the embedded NPs, pronouns, definite descriptions and proper names, as mentions (e.g. *it, the device, The HTC Touch Diamond*). This was done regardless whether they were linked to the topic or not, and was

achieved using PAlinkA [4]. Our corpus contains a total of 3372 markables. The second step was to mark links between these markables.

On the basis of corpus investigation, we decided to focus on 4 types of relations that are useful for our IQA task.

2.2. **Coreference.** This corresponds to the classical notion of coreference as defined by [3]. This is the most frequent relation and is transitive forming coreferential chains. Simple coreference should be carefully distinguished from relations SET OF and SIBLINGS (presented below), as sometimes the distinction between them is not straightforward.

The COREFERENCE relation is marked only between markables that refer to the same entity in the real world. This includes coreferential links such as identity, synonymy, generalization and specialization, but they were not explicitly distinguished as proposed in [1]. In general, only definite descriptions that stand in the relationship of identity (same head: *a smart phone - The Touch Pro smartphone* ; pronouns: *Opera Mobile - it*) or synonymy (*the device - the phone*) with the antecedent were marked as coreferential. Usually an anaphoric expression is linked to the previous mention of the NP in the document, but it can be also linked to the first mention.

Text in brackets and text between dashes after an NP is marked as coreferential with this NP (as long as it definitely refers to the NP): e.g. *[the XV6800 ([Verizon Wireless]) variant of [the device]]*. For this type of coreferential link, the anaphor should be linked back to the nearest antecedent.

2.3. **Set of.** One characteristic of the Wikipedia pages discussing products is that they can describe several versions of the same product. This is normally marked by adding a prefix or suffix to the original name. Given the purpose of this research, such links should be identified in texts, but they should not be marked as identity as they refer to entities with different characteristics. For this reason, we add a SET OF link from the markable to the antecedent that describes the set (i.e. the topic of the article). E.g. A modified version of *[the Hero]*, *[the HTC Droid Eris]*, was released on the Verizon Wireless network on November 6, 2009.

The SET OF relation is used to link members of hyperonymy hierarchy: it links less general markable to more general one. For our corpus, this happens when a phone has several submodels. The link is always added from the submodel to the nearest markable that corefers to the topic. SET OF is also used to identify more general categories than the topic as it happens to markables in copular relation like in the following example: *[The HTC Dream] is [an Internet-enabled 3G smartphone]*. In this case the relation will be from *"The HTC Dream"* to *"an Internet -enabled 3G smartphone"*. This gives the possibility to collect more information about the topic.

2.4. **Alias.** Another characteristic of Wikipedia product articles is that the same product can be referred to using different names. This is a special case of coreference relation where a completely different name is used for the product and not a substring of the original name. This relation is usually indicated by phrases such as *is also named as* and *has codename*. E.g. *[The HTC Touch Diamond]*, also known as *[the HTC P3700]* or *[its]* codename *[the HTC Diamond]*, is a ...

Relation ALIAS is quite straightforward and is used to indicate situations when different names are used for the same entity. This relation is quite common in our corpus and usually is introduced by a limited set of verbal phrases. The link is always from the markable that represents the alias to the nearest markable that corefers with the topic.

2.5. **Siblings.** For interactive question answering it is very important to identify when two entities differ in terms of only a few characteristics. This is due to the fact that in case of ambiguity a user should be presented with close alternatives and be asked to decide between them. This relation happens when the two entities are in a SET OF relations with the topic of the article. We call the link between these entities SIBLINGS relation to indicate the near-identity between them. In our corpus this phenomenon happens quite often when the same mobile phone is distributed by different operators with slightly different features, and possibly with a different name. This relation is not explicitly marked during the annotation process, but it can be inferred on the basis of the above annotation.

In our corpus we annotated a total of 668 coreferential relations, 83 SET OF relations and 59 ALIAS relations.

3. Experiment

The corpus annotation described in the previous section revealed some regularities in the way expressions refer to the topic which could be captured using a rule-based approach. This next section briefly presents these rules followed by preliminary evaluation results.

3.1. **Rule-based coreference resolution method.** The rule-based method developed here relies on high precision rules that use particularities of the documents to be processed, with emphasis on product names. Different rules are used to target the different types of relations described above. Given that our current focus is on the identification of expressions that refer or are linked to the main topic of the article, we rely on the markables annotated by humans. This allows us to ensure that no errors are introduced in the process as a result of wrongly identified markables.

The identification of all the relations is combined into a pipeline, where already identified relations are used for further processing. First, ALIAS relations are found and alternative names of the topic are added to the list. This helps to reveal all possible ways the topic can be referred to throughout the text. Given the fact that we are interested not only in tracking the topic but also all subtopics, the next step is identification of SET OF relations. This stage yields a list of subtopics and at a later stage they are treated in a similar way to topic expressions in order to identify all coreference chains. The last step is discovering all coreference links for topic and subtopics.

The following list shows a few examples of rules used here:

- A markable corefers with the topic if the topic ends with the markable after determiners are removed e.g. the markable *the Bold 9700* corefers with the topic *The Blackberry Bold 9700*
- Expressions such as *also called, formerly known as* between two markables indicate that the second markable is an alias for the first
- If the topic is included in a longer markable, the relation between the markable and the topic is SET OF e.g. the markable *The GSM BlackBerry Storm* is in the relation of SET OF with the topic *The BlackBerry Storm*

3.2. **Evaluation.** Evaluation of the rule-based approach presented above revealed several issues that need to be addressed. We used the MUC score [7] to assess the accuracy of the topic identification.

As it was mentioned above the main assumption of our research is that Wikipedia articles describe the topic and provide more information about it. Therefore as a baseline all subjects in the corpus were annotated as coreferential with the topic. Connexor's Machinese [1] was employed for annotation of the corpus with syntactic relations and then tag SUBJ was used to identify all subjects in the text.

Evaluation of the system output showed that it can identify the topic with an accuracy of 75.33% f-measure, where as the baseline achieves only 14.07% f-measure.

During the development of our method several issues that affect the performance of the system were identified. First inconsistencies in the annotation of the gold standard were identified and corrected. This issue was addressed by correcting the annotation of the files.

Another problem was caused by the contents of some articles which did not describe a model of a phone but the whole series of phones. In this case, the article does not have a main topic, but rather many subtopics. Given the

[1] http://www.connexor.eu/technology/machinese/

fact that our experiment assumed the presence of the main topic, this kind of texts were not processed correctly.

Automatic processing of the texts relies on the peculiarities we identified while studying the organisation of Wikipedia articles, e.g. it was noticed that the first markable in the files denotes the topic. However this rule had exceptions and so the output of the system was incorrect in some cases.

4. Conclusions

This extended abstract has presented a rule-based method for topic tracking in Wikipedia articles. The results of the algorithm are promising for most of the texts as it relies on the presence of a regular structure in the articles. Investigation of the files for which the performance is rather low revealed that even humans have problems analysing them.

A conclusion of this research is that for our application, it is not possible to use the classical definition of coreference where entities are either coreferential or not. Instead, we need to define several near identity relations. As a result, it is not possible to apply the standard evaluation metrics directly. The full paper will discuss this issue as well.

References

[1] L. Hasler, C. Orăsan, and K. Naumann. NPs for Events: Experiments in Coreference Annotation. In *Proceedings LREC2006*, pages 1167 – 1172, Genoa, Italy, May, 24 – 26 2006.

[2] R. Mitkov. *Anaphora resolution*. Longman, 2002.

[3] V. Ng. Supervised noun phrase coreference research: The first fifteen years. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1396–1411, Uppsala, Sweden, July 2010.

[4] C. Orăsan. PALinkA: a highly customizable tool for discourse annotation. In *Proceedings of the 4th SIGdial Workshop on Discourse and Dialog*, pages 39 – 43, Sapporo, Japan, July, 5 -6 2003.

[5] M. Recasens, E. Hovy, and M. Antònia Martí. A typology of near-identity relations for coreference (nident). In *Proceedings of LREC 2010*, pages 149–156, Valletta, Malta, 2010.

[6] V. Stoyanov, C. Cardie, N. Gilbert, E. Riloff, D. Buttler, and D. Hysom. Coreference resolution with Reconcile. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 156–161, Uppsala, Sweden, July 2010.

[7] M. Vilain, J. Burger, J. Aberdeen, D. Connoly, and L. Hirschman. A model-theoretic coreference scoring scheme. In *Proceedings of the 6th Message Understanding Conference (MUC-6)*, pages 45 – 52, San Francisco, California, USA, 1995.

Research Institute in Information and Language Processing, University of Wolverhampton, United Kingdom

E-mail address: N.Konstantinova@wlv.ac.uk and C.Orasan@wlv.ac.uk