DE GRUYTER OPEN

STUDIES IN LOGIC, GRAMMAR AND RHETORIC 56 (69) 2018

DOI: 10.2478/slgr-2018-0050

Coreference Resolution for Anaphoric Pronouns in Texts on Medical Products

Jerzy Krawczuk[1], **Mariusz Ferenc**[2]

[1] Faculty of Computer Science, Bialystok University of Technology, Poland
[2] Ediom sp. z o.o., Poland

Abstract. Coreference resolution is the task of finding all expressions that refer to the same entity in a text. It is one of the higher level NLP (Natural Language Processing) tasks. It allows, for example, to extract more information about medical products from larger texts. A product such as 'ambidextrous gloves' may appear in a text in many different forms. For example, they could be referred to by the pronoun 'they', such as in this sentence. The algorithm presented in this paper finds pronouns and for each of them (except the pleonastic 'it') it creates a coreference candidate with entities that appeared earlier in the same sentence or in the previous sentence. Each candidate (pair of mentions) is described by 48 binary features which represent their grammatical and location properties. In the training set, each pair is marked as a coreference or not, based on which a decision tree classifier is trained. A classifier with a high precision of 0.94 and a decent recall of 0.61 were obtained on the training set, still with a good precision out of a sample of 0.64.

Introduction

Coreference resolution is the process of determining whether two expressions in natural language refer to the same entity in the real world. It is an important subtask in natural language processing systems. In particular, information extraction (IE) systems have revealed that coreference resolution is a very important component. In the sixth of a series of Message Understanding Conferences, a separate subtask was defined and evaluated (MUC-6 1995). This task involved the identification of coreference relation between noun phrases. Noun phrases are also considered in this paper, with the first noun phrase being of a common or proper type and the second noun phrase a pronoun. This condition was chosen due to the fact that such coreferences are the most common in texts concerning medical products, for example:

Stent grafts are larger stents used for larger arteries. They may also be made of a specialized fabric.

In the second sentence the pronoun *they* refers, or corefers, with the noun phrase *stent grafts*. If such coreference is discovered correctly, an information extraction system can also know that *Stent grafts may also be made of a specialized fabric*. Humans extract this information from a text easily, but it is still very challenging for the computer.

There are two main approaches in coreference resolution. The first one is deterministic (Lee et al., 2011) and the second one is based on machine learning. Machine learning approaches are mostly based on defining the coreference problem as a classification task (Soon et al., 2001). Specifically, a pair of NPs is classified as coreferring or not based on rules learned from an annotated corpus. The deterministic approach defines specific rules based on, for example, parts of speech tags.

The rest of the paper is structured in the following way: in the *Related-works* section, several related works in the area of coreference resolution with a focus on statistical approach and pronouns are presented. In the *Methods* section, the used method is presented, starting from defining the mentions and coreference candidates; the so-called pleonastic 'it' (which is an 'it' not referring to another entity) is described, which is a challenging problem in itself. In the *Experiments* section, an experiment performed on scientific articles from PubMed, Wikipedia entries, and manufacturers' catalogs to learn a model and use it to find coreferred noun phrases for each pronoun is described. The last section, i.e. *Conclusions*, presents the conclusions and improvement ideas.

Related Works

The most influential article for this work was the statistical coreference system implemented in Stanford CoreNLP (Clark & Manning, 2015). It uses a mention pair model to predict whether or not a given pair is in coreference. The authors argue that pronouns often have only one clear antecedent, but their algorithm is not restricted solely to such coreferences. A logistic classifier is used to assign probability to each pair. As not only pronouns are considered, the authors use a second ranking algorithm to cluster the classified mentions. Five groups of features are used: distance, syntactic, semantic, rule-based, and lexical. Most of them are implemented in this work and described in detail in the next chapter.

In the previous work (Soon et al., 2001), mentions were called markables and the authors used only 12 features for pair candidates. These features were studied carefully to show which of them were the most important sources of errors. They discovered that 'prenominal modifier string match' was 42.1% for precision measure while 'inadequacy of current surface features' was 63.3% for recall. The results were achieved with the C5 decision tree learning algorithm, which is an updated version of the C4.5 (Quinlan, 2014) algorithm used in this paper. They reported a precision of 67.3% and a recall of 58.6% of their algorithm.

In an even earlier work, only the anaphora resolution for pronouns was described (Ge et al., 1998). The authors obtained a precision of 66% on 21 million words from Wall Street Journal texts, which is a result similar to the one achieved in this work. The most common pronouns presented there were referred to people, e.g. 'she' and 'he'. In this work, the most common pronouns are 'it' and 'they', which refer to products in the singular or in the plural, respectively.

A state-of-the-art deterministic approach was implemented by Staford NLP Group. It is the so-called fast rule-based coreference resolution (Lee et al., 2013), a multi-(ten)-sieve system that starts from higher precision rules and goes to lower precision but higher recall rules. Owing to the setting of appropriate weights on each sieve, the system enable balancing between precision and recall. The authors report results for precision and recall close to 70 percent.

There are also other approaches to learning coreferences and one of them is the mention-ranking models. In this approach, models score pairs of mentions for their likelihood of coreference, thus they operate in a simple setting where coreference decisions are made independently. Having independent actions enables an easy use reinforcement learning. Clark and Manning (2016) use these kinds of models in their research. They report a precision of 79% and a recall of 70% on CoNLL 2012 English Test Data. Apart from the standard features, such as distance between mentions candidates, string matching, and speaker identification, they use word embedding.

For some time, especially at the beginning of the development of anaphora resolution systems, there was considerable focus on rules, but the work mentioned in the previous paragraph shows that statistical approaches give better performance. Right now there is much less focus on them; however, there are still some advantages to such systems, e.g. rules that are easy to create and maintain as well as a more transparent error analysis.

Methods

C4.5 decision trees (Quinlan, 2014) were used to train the classifier that decided which noun phrases NP_i and NP_j ($j > i$) in a document are coreferent. In this section, the 3 main steps in the authors' algorithm are described, i.e.: mention detection (noun phrases), creation of a coreference candidate and the features describing it and, finally, building a classifier to distinguish coreferent mention pairs. The following section describes selected important aspects of the methods, which include assigning the number and gender to a noun phrase, as well as discovering the so-called pleonastic 'it', which does not refer to any previous entity mention in a text.

Mention detection is performed following part of speech tagging and parsing of sentences. Part of speech tagging tags each word in a sentence using a tag set; in the analyzed case, this is Penn (Santorini, 1990). Examples of parts of speech are:

- NN, NNS – noun in the singular and plural form,
- CC – conjunction, coordination,
- IN – conjunction, subordinating,
- JJ – adjective,
- DT – determiner,
- VBP – verb,
- PRP – pronoun.

After tagging words are grouped into sentence parts (phrases), for example:

- NP – noun phrase,
- VP – verb phrase,
- PP – prepositional phrase.

A parse tree (also called a syntax tree) is the result of such a process (Figure 1 and Figure 2). In this tree, the leaf nodes represent parts of speech whereas branch nodes represent parts of the sentence. The top (root) node S represents the whole sentence.

> *Ambidextrous gloves are extra soft and comfortable for every day use. They protect hands from a broad range of chemicals.*

In general, mentions in the form of noun phrases not containing other noun phrases are discovered. The mentions in the first sentence in Figure 1 are: [*Ambidextrous gloves*, *extra soft and comfortable*, *every day use*]. The mentions in the second sentence in Figure 2 are: [*They*, *hands*, *a broad range*, *chemicals*]

However, there are some exceptions to this rule. Inclusion of constructions such as NP(NP PP(IN NP))) was found useful. In this case,

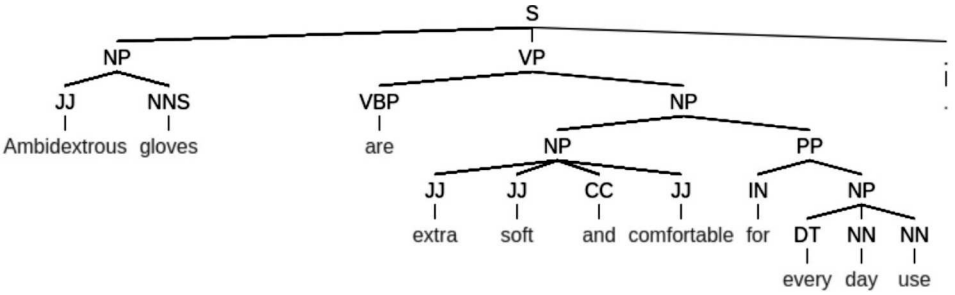

Figure 1. First sentence parsing tree

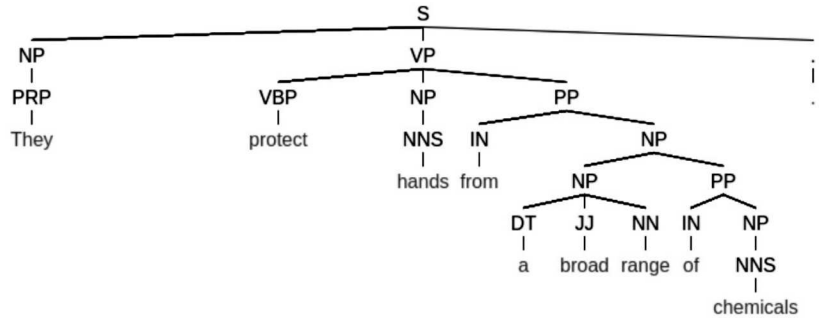

Figure 2. Second sentence parsing tree

the noun phrase can contain two other noun phrases separated by a preposition (IN), e.g.: *of*, *for*. Such noun phrases are present in both example sentences – it can be as long as *extra soft and comfortable for every day use* and its usefulness may not be obvious; in the second sentence, the usefulness of such a rule is more visible, i.e. *a broad range of chemicals*. This is the only exception that points to the inclusion of a more complex noun phrase; however, there are also certain other rules pointing to the exclusion of some of them:

Noun phrases containing a cardinal number (CD) only are not included, such as in the following sentence:

Ten is lower than eleven.

(S (NP (CD Ten)) (VP (VBZ is) (NP (QP (JJR lower) (IN than) (CD eleven)))))

Another example of filtered noun phrases is as follows:

Pair of jeans

(S (NP (NP (NNP Pair)) (PP (IN of) (NP (NNS jeans)))))

[*pair*] is not taken as a mention. Only [*jeans*] and [*pair of jeans*] are taken.

Furthermore, some instances of the so-called pleonastic 'it', described in detail in the section devoted to the pleonastic 'it', are not included, but a simple example would be:

It is raining.

The coreference candidate consists of two mentions (noun phrases). The first noun phrase is of a common, or proper, type while the second is always a pronoun. Both have to be in the same or in neighbouring sentences. In the previous two-sentence example, the coreference candidates are:

[*ambidextrous gloves ; they*]

[*extra soft and comfortable ; they*]

[*extra soft and comfortable for every day use ; they*]

[*every day use ; they*]

Each such pair (NP_j, NP_J) is called a coreference candidate. Only some of them are coreferent, while most of them are not. Such a pair is described by 48 binary features shown in Table 1. Twelve features describe the pair and eighteen describe each mention (noun phrase).

It was decided that the features would be binary. Hence, for example, there are two possible features, i.e. plural and singular, for the number feature. It is possible that no information is known about the number so both could be set to zero. A further implication is that there are two features for a number match, as it can be known that they: match, differ, or possibly one or two features are unknown, so there is no certainty. In the last case, both the '3: number match' and '4: number do not match' features are set to zero.

C4.5 decision trees (Quinlan, 2014) were used in addition to the implementation available in the suite of the machine learning software called Weka (Witten et al., 2016). The authors' early experiment show that this classifier performs best on the training set. Another considered algorithm, with slightly worse results, was logistic regression (Le Cessie & Van Houwelingen, 1992).

Finally, the ensemble approach is used on the testing test. The tree classifier is combined with a classifier based on the training set. Each feature vector from the testing set is first looked at in the training set. Since binary features are dealt with, many cases are expected to be described by the exact same features.

One of the biggest challenges in coreference resolution for anaphoric pronouns is to detect the so-called pleonastic 'it'. Pleonastic 'it' is the pro-

Table 1. Coreference candidate binary features

FEATURES FOR PAIR OF MENTIONS	
1	gender match
2	gender does not match
3	number match
4	number does not match
5	in same sentence
6	in next sentence
7	compatible
8	possible compatibility
9	at least one compatible mention between
10	at least two compatible mentions between
11	at least one possible compatible mention between
12	at least two possible compatible mentions between
FEATURES FOR EACH MENTION	
1–3	type (common, proper, pronoun)
4–5	number (plural, singular)
6–8	gender (male, female, neutral)
9	is starting sentence
10	demonstrative
11	is in verb phrase
12–16	parent phrase type (NP, VP, PP, ADJP, ADVP)
17	is left sibling VP
18	is right sibling VP

noun 'it' that does not refer to any previously mentioned entity in the text, for example:

It is raining.
It is important to note that 7–8 percent of patients implanted with standard monofocal lenses also notice glare and halos.

In both sentences, the pronoun 'it' that starts them does not refer to any previous mention. In general, there are two types of approach to discovering the pleonastic 'it': the rule-based approach and the machine learning approach.

In the discussed system, the rule-based system with 12 rules described in Lee et al. (2013) is implemented. The rules are written in the so-called tregex ("tree regular expressions") (Levy & Andrew, 2006), and part of such a rule takes the following form:

NP < (PRP = it) $.. (VP < ((/^V.*/</^(?:turns|turned)/)

The tregex checks if the sentence parse tree contains a noun phrase that contains the pronoun 'it' and if there is also a verb phrase that contains the verb *turns* or *turned* at the same tree level.

The tregex is not used directly; instead, the authors' own implementation operating directly on parse trees from Open NLP library (http://opennlp.sourceforge.net) is used. Open NLP is a project maintained by the Apache Software Foundation.

These rules work with high precision but there may be more pleonastic 'it's in the text, which would not be discovered by the rules; the authors' algorithm still may not assign a previously mentioned entity to them. It is possible that for some pronouns a coreferenced entity will not be assigned due to the fact that the probability for each candidate will be too low (Table 2). A minimum threshold of 0.5 is set to consider candidate mention pairs as those coreferenced. If for some 'it's in the text the algorithm does not assign a probability higher than 0.5 for any previous mention, it may be said that it is comparable to discovering a pleonastic 'it' using the machine learning approach. Much depends on the learning set and the number of occurrences of such pronouns. In the analyzed test set, such cases did occur, for example in the following sentence:

> *It's quite possible the Max flow meter you are using has been in place and properly functioning for 15–20 or more years. This means that when it comes time to replace your system the exact model you have has probably been updated.*

Table 2. Probabilities for coreference candidates for pleonastic 'it'

coreference candidate	probability
more years , it	0.001088
place , it	0.001088
the Max flow meter , it	0.001088

Not every 'it' that was not assigned was actually pleonastic, for example in the sentence *"AAC is no longer performed on children as the Ross procedure has superseded it"* coreference was not discovered.

Gender is mostly assigned by using dictionaries, but also using WordNET, a lexical database for the English language (Miller & Fellbaum, 1998). The step of algorithms for gender assignment are the following:

- if 'it' is a pronoun, check in dictionary, male:he,him,his,himself female:she,her,hers,herself neutral:it,its,itself,
- search in male and female name dictionaries; for neutral, search in country and company names,
- identify other companies' names by: inc, corp, llc, ltd,
- identify male or female by titles: mr, mister, mrs, ms, miss, missus,
- search male and female nouns in the dictionary, such as: boy-girl, dad-mum, father-mother...,
- if 'it' is a common noun look for hypernyms in WordNet dictionary; if 'it' is male or female, assign gender accordingly; for neutral look for: artefact, location, group, entity,
- otherwise gender is set to unknown.

The singular or plural number is discovered using the following algorithm:

- if 'it' is a pronoun, check in dictionary, singular: I, me, he, him, she, her, ... plural: we, us, our, they, them...,
- if 'it' is a part of speech, the tag is NNS or NNPS, use plural,
- if 'it' is a part of speech, the tag is NN or NNP, check prefix, singular: a, an, this, that plural: those, these,
- if the 'and' conjunction occurs, use plural, for example Tom and Jerry,
- otherwise unknown.

Experiments

The experiment was conducted by preparing a separate training and testing set. In the training set, all anaphoric pronouns were marked with a proper or a common noun, which corefer. It was used to build a classifier on which the test set was set and the results were checked manually.

The training set contained 30 documents, mostly PubMed articles, in addition to several Wikipedia entries, where 77 pronoun coreferences were discovered manually. The algorithm ran on this set, created from a total of 1541 coreference candidates (mentions pairs), with all 77 coreferencing pairs included in it (Table 3). Each pair was represented by 48 binary features. Based on this learning set, containing 1541 objects with 48 binary features, C4.5 decision tree classifier is learnt. The classifier classified 50 candidates as coreferent on the same training set, making only 3 mistakes,

Table 3. Confusion matrix on training set

classified	NO	YES	
1464	1461	3	**NO**
77	30	47	**YES**
	1491	50	**coreference**

achieving a very high precision = 47/50 = 0.94. Recall from this classifier on the training set was 47/77= 0.61.

The testing set contained 50 documents, PubMed articles as well as manufacturers' web pages describing their products. All coreferences were not manually marked there; instead, the assigned coreferences were analyzed using the authors' algorithm. This means that there could be more coreferences in the document. For example, out of the 84 coreferences discovered, 54 were correct, while 30 were wrong (Table 4). Moreover, there could be more anaphoric pronouns in the document that were not accounted for. This allowed to report precision = 54/84 = 0.64, but not recall.

For classification, not only the learned decision tree was used but also, for each candidate, the binary vector was compared with those in the learning set. If an exact same match was found, class was assigned based on the example from the learning set; if it was not found in the learning set, then the decision tree classifier was used. In 25 percent of cases an exact match was found in the learning set and assigned a class accordingly, while in the other 75 percent of cases, the decision tree classifier was used.

Table 4. Confusion matrix on the testing set, only for the discovered coreferences

classified	NO	YES	
358	328	30	**NO**
67	13	54	**YES**
	341	84	**coreference**

Conclusions

In the paper, the statistical approach to coreference resolution was introduced for anaphoric pronouns. The authors believe that such an approach can be helpful for automatically extracting more information about medial products from different types of natural language texts. 48 binary features describing pairs of mentions were created, i.e. the so-called coreference candidates. C4.5 decision tree classifier was used to learn high precision (0.94) rules on a training set. Since it was decided that binary representation would be used, it was expected that many pairs could be described by the exact same vectors, thus it was decided that the authors' training set would also be used as a classifier. First, it was checked whether the exact same vector occurred in the training set before the tree classifier was applied. This was found to work in 25 percent of cases in the testing set. Precision on the testing set was lower (0.64), but still encouraging and comparable to the results presented in literature, for example those achieved by Soon et al. (2001) (a precision of 0.67).

REFERENCES

Clark, K., & Manning, C. D. (2015). Entity-Centric Coreference Resolution with Model Stacking. In *Proceedings of the 53th Annual Meeting of the Association for Computational Linguistic and the 7th International Joint Conference on Natural Language Processing* (pp. 1405–1415). Association for Computational Linguistics, Beijing, China.

Clark, K., & Manning, C. D. (2016). Deep reinforcement learning for mention-ranking coreference models. In *Proceedings of the 2016 Conference on Empirical Methods on Natural Language Processing* (pp. 2256–2262). Association for Computational Linguistics, Austin, Texas.

Ge, N., Hale, J., & Charniak, E. (1998). A Statistical Approach to Anaphora Resolution. In *Proceedings of the Sixth Workshop on Very Large Corpora* (pp. 167–170). Association for Computational Linguistics, Montreal, Canada.

Le Cessie, S., & Van Houwelingen, J. C. (1992). Ridge estimators in logistic regression. *Applied Statistics*, *41*(1), 191–201. doi: 10.2307/2347628

Lee, H., Chang, A., Peirsman, Y., Chambers, N., Surdeanu, M., & Jurafsky, D. (2013). Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Computational Linguistics*, *39*(4), 885–916.

Lee, H., Peirsman, Y., Chang, A., Chambers, N., Surdeanu, M., & Jurafsky, D. (2011). Stanford's Multi-Pass Sieve Coreference Resolution System at the CoNLL-2011 Shared Task. In *Proceedings of the 15th Conference on*

Computational Natural Language Learning: Shared Task (pp. 28–34). Association for Computational Linguistics, Portland, Oregon, USA.

Levy, R., & Andrew, G. (2006). Tregex and Tsurgeon: tools for querying and manipulating tree data structures. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation* (pp. 2231–2234). European Language Resources Association, Genoa, Italy.

Miller, G., & Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database.* Cambridge: MIT Press.

Quinlan, J. R. (2014). *C4.5: Programs for Machine Learning.* Elsevier.

Santorini, B. (1990). *Part-of-Speech Tagging Guidelines for the Penn Treebank Project (3rd revision)* (Technical Report No. MS-CIS-90-47). University of Pennsylvania.

Soon, W. M., Ng, H. T., & Lim, D. C. Y. (2001). A Machine Learning Approach to Coreference Resolution of Noun Phrases. *Computational Linguistics*, *27*(4), 521–544.

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann.