Name Perplexity

Octavian Popescu

Abstract

The accuracy of a Cross Document Coreference system depends on the amount of context available, which is a parameter that varies greatly from corpora to corpora. This paper presents a statistical model for computing name perplexity classes. For each perplexity class, the prior probability of coreference is estimated. The amount of context required for coreference is controlled by the prior coreference probability. We show that the prior probability coreference is an important factor for maintaining a good balance between precision and recall for cross document coreference systems.

1 Introduction

The Person Cross Document Coreference (PCDC) task which requires that all and only the textual mentions of an entity of type Person be individuated in a large collection of text documents, is a challenging tasks for natural language processing systems (Grishman 1994). A PCDC system must be able to use the information existing in the corpus in order to assign to each person name mention (PNM) a piece of context relevant for coreference. In many cases, the contextual information relevant for coreference is very scarce or embedded in semantic and ontological deep inferences, which are difficult to program, anyway.

Unlike in other disambiguation tasks, like word sense disambiguation for instance, where the distribution of relevant contexts is mainly regulated by strong syntactic rules, in PCDC the relevance of contexts is a matter of interdependency. To exemplify, consider the name "John Smith" and an organization, say "U.N.". The context "works for U.N." is a relevant coreference context for "John Smith" if there is just one person named John Smith working for U.N.; if there are two or more John Smiths working for U.N., then "works for U.N." is no longer a relevant context for coreference. For the PCDC task, the relevance of the context depends to a great extent on the diversity of the corpus itself, rather than on the specific relationship that exists between "John Smith" and "works for U.N.".

Valid coreference can be realized when a large amount of information is available. However, the requirement that only contextually provable coreferences be realized is too strong; the required relevant context is not actually explicitly found in the text in at least 60% of the times (Popescu 2007).

This paper presents a statistical technique developed to give a PCDC system more information regarding the probability of a correct coreference, without performing deep semantic and ontological analyses. If a PCDC system knows that the prior probability for two PNMs to corefer is high, then the amount of contextual evidence required can be lowered and vice-versa. Our goal is to precisely define a statistical model in which the prior coreference probabilities can be computed and, consequently, to design a PCDC system that dynamically revises the context relevance accordingly.

We review the PCDC literature relevant for our purposes, present the statistical model and show the preliminary results. The paper ends with the Conclusion and Further Research section.

2 Related Work

In a classical paper (Bagga 1998), a PCDC system based on the vector space model (VSM) is proposed. While there are many advantages in representing the context as vectors on which a similarity function is applied, it has been shown that there are

Proceedings of NAACL HLT 2009: Short Papers, pages 153–156,
Boulder, Colorado, June 2009.

inherent limitations associated with the vectorial model (Popescu 2008). These problems, related to the density in the vectorial space (superposition) and to the discriminative power of the similarity power (masking), become visible as more cases are considered. (Gooi, 2004), testing the system on many names, empirically observes the variance in the results obtained by the same PCDC system. Indeed, considering just the sentence level context, which is a strong requirement for establishing coreference, a PCDC system obtains a good score for "John Smith". This is because the probability of coreference of any two "John Smith" mentions is low. But, as the relevant context is often outside the sentence containing the mention, for other types of names the same system is not accurate. If it considers, for instance, "Barack Obama", the same system obtains a very low recall, as the probability of any two "Barack Obama" mentions to corefer is very high. Without further adjustments, a vectorial model cannot resolve the problem of considering too much or too little contextual evidence in order to obtain a good precision for "John Smith" and simultaneously a good recall for "Barack Obama".

The relationship between the prior probabilities and the accuracy of a system is also empirically noted in (Pederson 2005). In their experiment, the authors note that having in the input of the system the correct number of persons carrying the same name is likely to hurt the results of a system based on bigrams. This happens because the amount of context is statically considered. The variance in the results obtained by a PCDC system has been noted also in (Lefever 2007, Popescu 2007).

In order to improve the performances of PCDC systems based on VSM, some authors have focused on methods that allow a better analysis of the context (Ng 2007) combined with a cascade clustering technique (Wei 2006), or have relied on advanced clustering techniques (Chen 2006).

The technique we present in the next section is complementary to these approaches. We propose a statistical model designed to offer to the PCDC systems information regarding the distribution of PNMs in the corpus. This information is used to reduce the contextual data variation and to attain a good balance between precision and recall.

3 Name Perplexity Classes

The amount of contextual information required for the coreference of two or more PNMs depends on several factors. Our working hypothesis is that we can compute a prior probability of coreference for each name and use this probability to control the amount of contextual evidence required. Let us recall the "John Smith" and "Barack Obama" example from the previous section. Both "John" and "Smith" are American common first and last names. The chance that many different persons carry this name is high. On the other hand, as both "Barack" and "Obama" are rare American first and last names respectively, almost surely many mentions of this name refer only to one person. The argument above does not depend on the context, but just on the prior estimation of the usage of those names. Computing an estimation of a name's frequency class, we may decrease or increase the amount of contextual evidence needed accordingly.

To each one-token name we associate the number of different tokens with which it forms a PNM in the corpus. For example, for "John" we can have the set "Smith", "F. Kennedy", "Travolta" etc. We call this number the perplexity of a one-token name. The perplexity gives a direct estimation of the ambiguity of a name in the corpus. In Table 1 we present the relationship between the number of occurrences (in intervals, in the first column) and the average perplexity (second column). The figures reported here, as well as those in the next Section, come from the investigation of the Adige500k, an Italian news corpus (Magnini 2006).

occurrences (interval)	average perplexity
1-5	4.13
6-20	8.34
21-100	17.44
101-1,000	68.54
1,000-5,000	683.95
5,000-31,091	478.23

Table 1. Average perplexity one-token names

We divide the class of one-token names in 5 categories according to their perplexity: very low, low, medium, high and very high. It is useful to keep separate the first and the last names. It has been shown that the average perplexity is three times lower for last names than for first names

(Popescu 2007). Therefore, the first and last names perplexities play different roles in establishing the prior probability of coreference. The perplexity class of two-token names is computed using the following heuristics: the perplexity class of two-token names is the average class of the perplexity of the one-token names composing it. If the perplexity classes of the one-token names are the same, then the perplexity of the whole name is one class less (if possible).

The perplexity classes represent a partition of the name population; each name belongs to one and only one class. In establishing the border between two consecutive perplexity classes, we want to maximize the confidence that inside each stratum the prior coreference probability has a low variance.

The relationship between the perplexity classes and the prior coreference probability is straightforward. The lower the perplexity, the greater the coreference probability, and, therefore, the lower the amount of relevant context required for coreference.

In order to decide the percentage of the name population that goes into each of the perplexity classes, we use a distributional free statistics method. In this way we can compute the confidence of the prior conference probability estimates.

We introduce two random variables: X, a random variable defined over the name population and Y, which represents the number of different persons carrying the same name. Let $X_1,\ldots,X_n$ be a random sample of names from one perplexity class, and let $Y_1,\ldots,Y_n$ be the corresponding values denoting the number of persons that carry the names $X_1,\ldots,X_n$. The indices have been chosen such that $Y_1\ldots,Y_n$ is an ordered statistics: $Y_1 \leq Y_2 \leq \ldots \leq Y_n$. Let F be the distribution function of Y. And let p be a given probability. If $F(Y_j)-F(Y_i) \geq p$, then at least 100p percent of the probability distribution is between Y_i and Y_j; it means that

$$\gamma = P[F(Y_j)-F(Y_i)] \geq p \quad (1)$$

is the probability that the interval (Y_i, Y_j) contains 100p percent of the Y values.

In our case, γ is the confidence of the estimation that 100p percent of names from a certain perplexity class have the expected prior coreference probability in a given interval.

The γ probability is computed with the formula:

$$\gamma = P(F(Y_j) - F(Y_i) < p) = 1-\int_0^p \Gamma(n+1)/(\Gamma(j-i))\,\Gamma(n-j+i+1)x^{j-i-1}(1-x)^{n-j+i}dx \quad (2)$$

where Γ is the extension of the factorial function, $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt$.

In practice, we start with an interval that represents the prior coreference probability desired for that perplexity class. For example we want to be γ = 80% sure that p = 90% of the two-token names in the "very low" perplexity class are names carried by a maximum of 2 persons. We choose a random sample of two-token names from that perplexity class, the size of the random sample being determined by γ and p – see equation (2). If the random sample satisfies (1) then we have the desired perplexity class. If not, the one-token names that have the highest perplexity and were considered "very low" are excluded – they are assigned to the next perplexity class - and the computation is re made.

In a preliminary experiment, using a sample of 25 two-token names from a part of the Adige500k corpus spanning two years, we have obtained the perplexity classes listed in Tables 2 and 3. In Adige 500k there are 106, 192 different one-token names, which combine into 429, 251 different two-token names and 36, 773 three-token names.

perplexity class	percentage
very high	5.3%
High	8.7%
Medium	20.9%
Low	27.6%
very low	37.5%

Table 2. First Name perplexity classes

perplexity class	percentage
very high	1.8%
High	3.36%
Medium	17.51%
Low	20.31%
very low	57.02%

Table 3. Last Name perplexity classes

The perplexity class of two-token names is computed as specified in the first paragraph of this page. In approximately 60% of the cases, a two-token name has a "low", or "very low" perplexity class. If a PCDC system computes the context

similarity based on words with special properties or on named entities, in general at least four similarities must be detected between two contexts in order to have a safe coreference. Our preliminary results show that coreferring on the basis of just one special word and one named entity for those names in "low" or "very low" does not lose more than 1,5% in precision, while it gains up to 40% in recall for these cases. On the other hand, for "very high" perplexity two-token names we were able to increase precision by requiring a stronger similarity between contexts.

The gain of using prior coreference probabilities determined by the perplexity classes is important, especially for those names that are situated at the extreme: "very low" perplexity with a big number of occurrences and "very high" with a small number of occurrences. These cases establish the interval for the amount of contextual similarity required for coreference.

However, the problematic cases remain when the perplexity class is "very high" and the number of occurrences is very big.

4 Conclusion and Further Research

We have presented a distributional free statistical method to design a name perplexity system, such that each perplexity class maximizes the number of names for which the prior coreference belongs to the same interval. This information helps the PCDC systems to lower/increase adequately the amount of contextual evidence required for coreference.

In our preliminary experiment we have observed that we can adequately reduce the amount of contextual evidence required for the coreference of "low" and "very low" perplexity class. For the top perplexity class names the requirement for extra contextual evidence has increased the precision.

The approach presented here is effective in dealing with the problems raised by using a similarity metrics on contextual vectors. It gives a direct way of identifying the most problematic cases for coreference. Solving these cases represents our first objective for the future.

We plan to increase the number of cases considered in the sample required to delimit the perplexity classes. The equation (2) may be developed further in order to obtain exactly the number of required cases for each perplexity class.

References

A. Bagga, B. Baldwin.1998. *Entity-based Cross-Document Co-referencing using the Vector Space Model*, In Proceedings ACL.

J. Chen, D. Ji, C. Tan, Z. Niu.2006. *Unsupervised Relation Disambiguation Using Spectral Clustering,* In Proceedings of COLING

C. Gooi, J. Allan.2004. *Cross-Document Coreference on a Large Scale Corpus*, in Proceeding ACL.

R. Grishman.1994. *Whither Written Language Evaluation?* In proceedings Human Language Technology Workshop, 120-125. San Mateor.

E. Lefever, V. Hoste, F. Timur.2007. *AUG: A Combined Classification and Clustering Approach for Web People Disambiguation*, In Proceedings of SemEval

B. Magnini, M. Speranza, M. Negri, L. Romano, R. Sprugnoli. 2006.I-CAB – the Italian Content Annotation Bank. LREC 2006

V., Ng.2007. *Shallow Semantics for Coreference Resolution,* In Proceedings of IJCAI

T. Pedersen, A. Purandare, A. Kulkarni. 2005. *Name Discrimination by Clustering Similar Contexts*, in Proceeding of CICLING

O. Popescu, C. Girardi, 2008, *Improving Cross Document Coreference*, in Proceedings of JADT

O. Popescu, B. Magnini.2007, *Inferring Coreference among Person Names in a Large Corpus of News Collection*, in Proceedings of AIIA

O. Popescu, B. Magnini.2007. *Irst-bp: WePS using Named Entities*, In Proceedings of SEMEVAL

O. Popescu, M. Magnini, L. Serafini, A. Tamilin, M. Speranza.2006. *From Mention to Ontology: a Pilot Study*, in Proceedings of SWAP

Y. Wei, M. Lin, H. Chen.2006. *Name Disambiguation in Person Information Mining,* In Proceedings of IEEE