Person Cross Document Coreference with Name Perplexity Estimates

Octavian Popescu
popescu@racai.ro

.

Abstract

The Person Cross Document Coreference systems depend on the context for making decisions on the possible coreferences between person name mentions. The amount of context required is a parameter that varies from corpora to corpora, which makes it difficult for usual disambiguation methods. In this paper we show that the amount of context required can be dynamically controlled on the basis of the prior probabilities of coreference and we present a new statistical model for the computation of these probabilities. The experiment we carried on a news corpus proves that the prior probabilities of coreference are an important factor for maintaining a good balance between precision and recall for cross document coreference systems.

1 Introduction

The Person Cross Document Coreference (Grishman 1994) 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 one of the challenging tasks for natural language processing systems. In the most general case the corpus itself is the only available source of information regarding the persons mentioned and we consider that this is the case in this paper. A PCDC system must be able to use the information existing in the corpus in order to assign to each personal name mention (PNM) a piece of context. The coreference of any two PNMs is decided mainly on the basis of the similarity of the pieces of contexts associated with them. A successful PCDC must accurately extract the relevant context for coreference.

However, the context relevance is not absolute. Whether the contextual information uniquely individuates a person is a matter of probability. This paper presents a statistical technique developed to provide a PCDC system with more information regarding the probability of a correct coreference. The reason for developing this technique is twofold: (i) the relevant coreference context depends on the corpus itself and (ii) valid coreferences require a large amount of information, which is unavailable in the majority of cases.

The first reason is linked to a particularity of the CDC task that makes it more complex than other NLP tasks. Unlike in other disambiguation tasks, in the CDC tasks the relevant coreference context depends on the corpus itself. In word sense disambiguation, for instance, the distribution of the relevant context is mainly regulated by strong syntactic and semantic rules. The existence of such rules makes it possible for the disambiguation decisions to be made considering the local context. On the other hand, the distribution of the PNMs in a corpus is rather random and the relevant coreference context is a dynamic variable depending on the diversity of the corpus, that is, on how many different persons with the same name share a similar context. To exemplify, consider the name “John Smith” and an organization, say “U.N.”. The extent to which “works for U.N.” in “John Smith works for U.N.” is a relevant coreference context depends on the diversity of the corpus itself. If in that corpus, among all the “John Smiths” there is only one person who works for “U.N.” then “works for U.N.” is a relevant coreference context, but if there are many “John Smiths” working for U.N., then “works for U.N.” is not a relevant coreference system; in this last case, more contextual evidence is needed in order to correctly corefer the “John Smith” PNMs. The relevance of a context for coreference also depends on the corpus, not only on the specific relationship that exists between “John Smith” and

Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pages 997–1006,
Singapore, 6-7 August 2009. ©2009 ACL and AFNLP

"works for U.N.". Thus, A PCDC system must have access to global information regarding the PNMs.

The second reason comes from practical considerations. The amount of information required to correctly infer PNMs coreferences is not present in corpus in a computationally friendly way. In many cases the relevant coreference information is embedded in semantic and ontological deep inferences, which are difficult to program In as much as 60% of the cases, two documents containing the same name, from a news corpus, lack contexts which are directly similar and big enough to correctly decide on the coreference.

We propose a new method to control the amount of contextual coreference required for correct coreferences. Rather than having fixed rules deciding on the size of the context surrounding a PNM, we propose a probabilistic approach that requires contextual evidence for coreference differentially, by considering the prior probability of the coreference of two PNMs; the higher this probability is, the less their correct coreference depends on the context and vice versa. We present a statistical model where the prior coreference probabilities are computed considering only the corpus itself, and we show how these probabilities are used by a PCDC system that dynamically revises the amount of context relevant for coreference.

In Section 2 we review the CDC relevant literature. In section 3 we analyze the data from annotated coreference corpora and we individuate a specific problem, setting up a working hypothesis. In Section 4 we develop a statistical model for computing the prior coreference probabilities and in Section 5 we present the results obtained by applying it to a large news corpus. In section 6 a direct evaluation on CDC is carried on a test corpus. In Section 7 we show how the proposed techniques extends naturally to a strategy of construction relevant test corpora for CDC task. The paper ends with the Conclusion and the Future Research section.

2 Related Work

In a classical paper (Bagga and Baldwin 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 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.

Testing the system on many names, (Gooi and Allan, 2004), it has been noted empirically that the accuracy of the results varies significantly from name to name. 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 happens because the prior probability of coreference of any two "John Smiths" mentions is low, as this is a very common name and none of the "John Smith" has an overwhelming number of mentions. But 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 and the relevant coreference context is found very often beyond the sentence level. 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".

In an experiment using bigrams (Pederson et al. 2005) on a news corpus, it has been observed that the relationship between the amount of information given to a PCDC system and the performances is not linear. If the system has received in input the correct number of persons with the same name, the accuracy of the system has dropped. A typical case for this situation is when there is a person that is very often mentioned, and few other persons having few mentions; when the number of clusters is passed in the input, the clusters representing the persons who are rarely mentioned are wrongly enriched. However, this situation can be avoided if there is a measure of how probable it is to have a certain number of different persons with the same name, each being mentioned very often in a newspaper.

Recently, there has been a major interest in the PCDC systems, and, in the last two years, three important evaluation campaigns have been organized: Web People Search-1 (Artiles et al. 2007), ACE 2008 (www.nist.gov/speech/tests/ace/). It has been noted that the data variance between training and test is very high (Lefever 2007). Rather than being a particularity of those corpora, the problem is general. The performances of a bag of words VSM depends to a very high extent on the corpus diversity (see Section 3).

For reliable results, a PCDC system must have access to global information regarding the coreference space.

Rich biographic facts have been shown to improve the accuracy of PCDC (Mann and Yarowsky 2003). Indeed, when available, the birth date, the occupation etc. represent a relevant coreference context because the probability that two different persons have the same name, the same birth date and the same occupation is negligible. However, it is equally unlikely to find this information in a news corpus a sufficient number of times. Even for a web corpus, where the amount of this kind of information is higher than in a news corpus, the extended biographic facts, including e-mail address, phones, etc., contribute only with approximately 3% to the total number of coreferences (Elmacioglu et al. 2007).

In order to improve the performances of the PCDC systems based on VSM, some authors have focused on methods that allow a better analysis of the context by extracting the dependency chains (Ng 2007). The special importance of pieces of context has been exploited by implementing a cascade clustering technique (Wei 2006). Other authors have relied on advanced clustering techniques (among others Han et al. 2005, Chen 2006). However, these techniques rely on the precise analysis of the context, which is a time consuming process. It has been also noted that, in spite of deep analysis, the relevant coreference context is hard to find (Vu 2007).

The technique we present in the next sections 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 Data Analysis

In this Section we present the data analysis of the PNMs. We are interested in establishing a relationship between the distribution of the PNMs and the relevant context for coreference. As mentioned in the preceding sections, the amount of the relevant context for coreference cannot be decided prior to the investigation of that particular corpus. The performances of a bag of words VSM with a prior defined context approach will vary greatly from corpus to corpus. We have run the following experiment: we have considered the training and test corpora used in Web People Search-1 (WePS-1), which are web page corpora, and we have implemented a bag of word approach with two variants of clustering: agglomerative (A), and hierarchic (H). We have randomly chosen a set of seven names from training and test (14 names in total) and we have compared the results applying the two systems, A and H, on each set of names. In Figure 1 we present the results obtained. The figures on the vertical axes are computed using $F_{\alpha=0.5}$ formula.

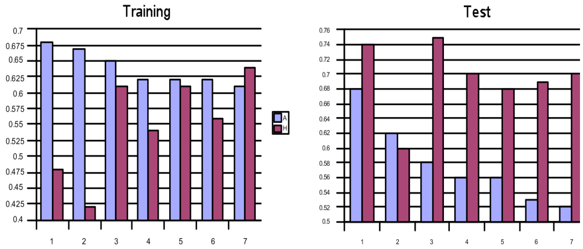


Figure 1. Variation between training and test

We have noticed a great variation in the behavior of the two systems. In order to search for an explanation for this difference we have looked at the distribution in the two corpora of the Named Entities, of the words denoting professions and of the meta-contextual information - e-mails, urls, phones, and addresses. It turns out that these types of contextual information are distributed between training and test approximately evenly. (see Table 1a,b).

Profession	training occ.	test occ.
Doctor	543	668
Lawyer	277	385
Professor	523	490
Researcher	340	166
Teacher	617	569
Coach	467	471
Actor	998	790

Table 1a. Profession words in training and test

Address	training occ.	test occ.
Phone	1,109	1,169
Fax	606	426
e-mail	3,134	2,186

Table 1b. Meta-Context in training and test

By manually investigating the training and test set of our experiment we have reached the conclusion that the reason for the difference is two fold: firstly, while the distribution of the words denoting profession is similar, in the test set the modifiers, for example “internist”, “neurosurgeon” for “doctor”, are more frequent. Secondly, the number of different persons having the same name is, on average, higher in test than in training. The results plotted in Figure 1 show that it is not a question of which algorithm is better, but rather that there are different cases where one approach is preferred over the other. The problem we face is deciding when it is appropriate to use one or the other.

To induce from the corpus itself when a piece of context is or is not a relevant context requires deep ontological inferences and a very powerful tool of semantic analysis of the context. Consider for example two words denoting profession, “doctor” and “researcher”, and their possible modifiers “internist”, “neurosurgeon”, and “professor” and “PhD”. In the first case it is certain that the coreference is not possible, while in the second the coreference is very probable. To find out such relationships is computationally very hard. However, the analysis carried out further shows that we can avoid making such computations in most of the cases.

The number of different persons is a parameter that cannot be known beforehand. However, not all the names behave alike with respect to coreference. There are noticeable differences between names; for example less than 5 000 first names cover approximately 96% of the total of first names, while for the same percentage of coverage more than 70 000 of last names must be considered (Popescu et al. 2007). Let us call perplexity of a name the number of different persons that carry it. The search space depends directly on the name perplexity. The bigger the perplexity, the larger the amount of information required for the correct coreference must be. It seems natural that the amount of contextual evidence required by a PCDC depends on the name perplexity.

In order to evaluate the relationships between the context and the name perplexity, we need an annotated corpus. We have used the I-CAB corpus (Magnini et al. 2006), which is a four-day news corpus fully annotated, coreference relationships included. The documents in this corpus are entire pieces of news. For each PNM we have counted how many contexts containing specific information about the person carrying the respective name is present in that particular document. There are many types of contexts that refer to a person, but some of these types are very infrequent. We considered only those types of information that are present at least 5% of the times in the context surrounding a PNM. Table 2 presents the results of this investigation.

	occ.	diff occ	entities
First Names	2299	676	1592
Last Names	4173	1906	2191
Middle Name	110	44	41
Activity	973	322	569
Affiliation	566	399	409
Role	531	211	317
Family Relation	133	46	94

Table 2. Name perplexity and context

On the second column the total number of occurrences is listed, on the third column how many of these occurrences have different values (no case sensitive string match), and on the fourth column the number of different persons (Entities) having that information. The entries “activity”, “affiliation”, and “role” represent pieces of context where the respective information is directly expressed (no inferences). We call this type of context professional context and for approximately 30% of the PNMs, one of the above three types of professional contexts is present.

The perplexity of the first names, computed as the ratio between the fourth column and the third column is two times bigger than the perplexity of the last names. The lowest name perplexity is obtained by the names having a middle name - a name with at least three tokens – and it is very close to 1 (1.07). Comparatively, the highest perplexity of two tokens name is 3. The relationship between the number of tokens of a name and its perplexity is straightforward: for names with more than four tokens the perplexity is 1 in 99,6% of the cases (the name by itself is a relevant context for coreference).

In approximately 74% of the cases there is just one entity corresponding to a two-token name. Considering any two PNMs of the same name the similarity of two of the professional contexts guarantees the correct coreference. However, two professional contexts are present in only 4% of the cases. There are just four cases when considering just one professional attribute was misleading, and all these cases are high perplexity names. Moreover, in the case of many low perplexity names, the contexts could be minimally similar in order to correctly corefer any two PNMs of that respective name.

This analysis shows that there is a direct relationship between the name perplexity and the relevant coreference context. However, the average figures are not very informative, as the variance of perplexity is very high. Rather than focusing on the exact figure for name perplexity, we will try to partition the names according to their perplexity and to link each partition to a specific behavior with respect to coreference. The partitioning technique should ensure that the variance of the name perplexity within the same partition is low and that a specific amount of context should lead to the correct coreference decision for the great majority of names within that partition.

Our working hypothesis is that we can estimate the name perplexity within each partition and use this information 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. Having an estimation of a name's perplexity, we may decrease/increase the amount of contextual evidence needed.

4 (p, γ) Statistical Model

Let $\mathcal{D}$ be the set of all PNMs from a given corpus $\mathcal{C}$ and let $\mathcal{D}_N$ be the set of corresponding names. We want to find a partition $\mathcal{P}$ of $\mathcal{D}_N$ such that within each partition the name perplexity varies only within predicted margins. Let X be a random variable with uniform distribution over $\mathcal{D}_N$ and let Y be the random variable defined by X's name perplexity. Let us suppose that we want $\mathcal{P} = \{p_1, p_2, \ldots, p_m\}$ to be a partition of $\mathcal{D}_N$, where the percentage of each partition class is p_i: the first partition class contains p_1 percentage of the name population, the second partition class contains p_2 percentage of the name population and the last partition class contains $p_m = 1 - \Sigma p_i$ percentage of the name population.

If we knew the distribution function of Y, let's call it F, we would simply determine ξ_i from equation 1, where $P_i = \Sigma p_k$, $k \leq i$:

$$\xi_i = F^{-1}(P_i) \Leftrightarrow F(Y<\xi_i) = P_i \qquad (1)$$

and we would know that in each partition p_i the name perplexity is between ξ_{i-1} and $\xi_{i,}$ with $\xi_0 = 0$. However we do not know F. Fortunately, we can estimate ξ_i.

There is no restriction that may impose a particular form for F; for example, the normal distribution hypothesis of name perplexity is ruled out by a χ^2 test with 96.5% confidence for the 14 names chosen from WePS-1 (see Section 3, first paragraph).

We are going to present a distributional free method for constructing the partition $\mathcal{P}$. The advantage of this method is that it does not depend on any assumption about the PNMs distribution.

Let us consider $X_1, X_2, \ldots, X_n$ a sample of independent and identical distributed names from $\mathcal{D}_{N.}$ By rearranging the indexes, without losing the generality, let us suppose that $Y_1, Y_2, \ldots, Y_n$ is ordered, that is $Y_1 \leq Y_2 \leq \ldots \leq Y_n$. Even if we do not know what form F has, we can still use equation (1) in order to estimate ξ_i. The expected value of $F(Y_i)$ is (Hogg, Mckean, Craig 2006):

$$E[F(Y_i)] = i/(n+1) \qquad (2)$$

which is an estimation of how much mass probability is on the left of Y_i. In our terms, we estimate that $E[F(Y_i)]$ percentage of the name population has a name perplexity lower than Y_i.

For a given number ξ, the percentage of the name population having the name perplexity at most ξ is determined by finding the smallest Y_i greater than ξ and use the equation (2) to estimate $E[F(Y_i)]$.

In order to build the partition $\mathcal{P}$ we are interested in the percentage of the name population that has the perplexity between two given values. Let (Y_i, Y_j) be the smallest interval that includes these two values. We can estimate the percentage of the name population that has a perplexity between Y_i and Y_j. This estimate is simply $F(Y_j) - F(Y_i)$. We can use directly equation (2) to estimate this difference. However, it is more important to have a confidence interval for this estimate, that is we want to know what the probability is that the interval (F(Yi), F(Yj)) contains at least a given percentage of the population, p. The optimal partition $\mathcal{P}$ is the one that maximizes the confidence in the fact that within each of the partition classes as many names as possible have the name perplexity in a given interval.

Let p be a given real number between (0,1) representing the mass probability that goes into the interval $(F(Y_i), F(Y_j))$. Let $\gamma = P(F(Y_j) - F(Y_i) \geq p)$. Fortunately γ has a distribution that does not depend on F. More precisely, γ has a beta distribution given by the function in formula (3):

$$\gamma = P(F(Y_j) - F(Y_i) \geq p) =$$

$$\int_p^1 \Gamma(n+1)/(\Gamma(j-i))\ \Gamma(n-j+i+1)x^{j-i-1}(1-x)^{n-j+i}dx \quad (3)$$

The Γ, called the gamma function, is the extension of the factorial, $\Gamma(x) = \int_0^{\infty} t^{x-1}e^{-t}dt$. The gamma function has the property that $\Gamma(x) = \Gamma(x-1)\ \Gamma(x-2)\dots\ \Gamma(1)$; for x an integer, as the arguments in the formula (3) are, $\Gamma(x) = (x-1)!$

The formula (3) gives us a method of building the partition $\mathcal{P}$. Let us start with a set of perplexity intervals: $(\xi_0, \xi_1]$, $(\xi_1,\xi_2]$, $\dots(\xi_{m-1}, \xi_m]$. We partition the names in $\mathcal{D}_N$ such that we maximize the confidence γ that at least p_i percentage of the name population has a name perplexity in $(\xi_{i-1}, \xi_i]$. We chose an independent and identical distributed sample X of names to which the ordered sample Y of name perplexity values corresponds. We start with the lowest perplexity interval and determine p_1,γ_1 and Y_0, Y_{i1}, such that $Y_0 \leq \xi_0 \leq \xi_1 \leq Y_{i1}$ and $\gamma_1 = P(F(Y_{i1}) - F(Y_0) \geq p_1)$. The ith index varies according to the desired $\gamma_{1,}$ when p_1 is given, and vice-versa. We can choose i1 with m-1 liberty grades. Once we are satisfied with the values (p_1,γ_1), we search for the i2th index such that $Y_{i1} \leq \xi_1 \leq \xi_2 \leq Y_{i2}$ and (p_{i2},γ_{i2}) have the desired value. The process continues till the penultimate (p_{m-1},γ_{m-1}). We have no liberty in choosing the (p_m,γ_m).

We can compute the size of the sample needed for guaranteeing a minimum γ and p.

Let us give an example. Suppose that $(\xi_0, \xi_1] = (0,2]$. Thus we are interested in finding p, the percentage of the name population such that we can be γ sure that at least p names have a perplexity between 1 and 2 inclusive. We take a random sample of n = 30 and suppose the smallest index i1 such that $Y_i \geq 3$ for all i > i1 is 17 . We want to compute the confidence γ that at least p = 60% of the name population has the name perplexity within (0,2]:

$$\gamma = 1 - \int_0^{0.6} 30!/(16!15!)x^{15}(1-x^{14})dx =$$

$$= 1 - k(\int_0^{0.6} x^{15}dx + \int_0^{0.6} x^{29}dx) =$$

$$= 1 - k[(1/16)(6/10)^{16} + (1/30)(6/10)^{30}]$$

$$\geq .965$$

In practice we want to have optimal values for p and γ; a large p implies a small γ and vice-versa. The optimality is determined by the accuracy of the CDC system: we want to have the largest possible percentage of names into each partition such that our confidence that the names inside each partition have the same perplexity.

It is useful to work the equation (1) backwards. Suppose that we established the first partition class of $\mathcal{P}$ - we have found the i1th index, p_1, and γ_1. Now we refer only to the names in the partition class. We can compute the probability that a certain percentage of the names *within that particular partition class* have a given name perplexity. That is, we consider a random *sample inside the partition class,* X, and its correspondent random variable Y, as above. The confidence that $p_{1inside}$ percentage of names have the name perplexity ξ_p within the interval (Y_0, Y_{ith}) is:

$$P(Y_0 < \xi_p < Y_{i1}) = \Sigma_k\ ({}_k{}^n)\ p^k(1-p)^{n-k} \qquad (4)$$

$({}_k{}^n)$ represents the k-combinations of size n.

By taking advantage of the bootstrapping method (Efron and Tibshirani 1993) we do not have to resample inside the partition class. We use the Y_0, .., Y_{i1} values with replacement. Using (4) we obtain $p_{1inside}$ which shows us which percentage inside the partition class has the name perplexity within $(Y_0, Y_{i1}]$. And consequently we can compute $\gamma_{1inside}$. Finally we are able to formulate the following statement about each partition class:

> In the ith partition class enter p_i percentage of the name population with a confidence γ_i. Inside this partition class we are $\gamma_{iinside}$ confident that $p_{iinside}$ percentage of the names have a name perplexity within $(Y_{i1-1}, Y_{i1}]$.

The p and γ indicate the theoretical values that define the partition. In practice the exact distribution of the names into the subset is unknown, therefore each heuristics that computes the perplexity creates its own distribution. The values $\gamma_{iinside}$ and $p_{iinside}$ control how much a certain heuristics departs from the theoretical values. The optimal heuristics have very big figures for $\gamma_{iinside}$ and $p_{iinside}$.

In the next section we present an experiment carried on a news corpus. We show how the above model leads to a stable partition of names and that inside each partition class reliable (p,γ) values can be computed.

5 Name Perplexity Partition

For the experiment described in this section, we have used a two-year part of the seven-years Italian local newspaper corpus called Adige500k Corpus (Magnini 2006).

We describe below how we compute the perplexity class for the one-token names and two-tokens names respectively. As mentioned in section 3, the name perplexity decreases rapidly for tree-token or more names. If desired, the same technique could also be applied for those names. In Adige500k there are 106, 187 different one-token names; 429, 243 two-token names; 36, 773 three-token names; 5, 152 four-token names, 940 four token names and less than 300 different four-token or more names.

An estimate of the name perplexity of the one-token names is the size of the different one-token names with which it forms a complete PNM in the corpus. For example for the first name "John" the estimation of its perplexity is the size of the one-token last names it combines with in forming PNMs, like "Smith, Travolta, Kennedy" etc. The bigger the size of its complementary names, the higher is its name perplexity. In Table 3 we present the figures of these estimates.

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 3. Average perplexity one-token names

We start with a five name perplexity classes: "very low" (VL) , "low" (L) , "medium", (M) "high" (H) and "very high" (VH). The name perplexity of a two-token name is interpolated from the name perplexity of its components. We used the following heuristics: the name perplexity class is the average name perplexity classes of its one-token name. If the name perplexity classes are the same then the name perplexity class of the whole name is one class less (if possible).

In order to compute the borderline between two consecutive classes we apply the (p, γ) method. We selected 25 two-tokens names and we manually investigate their occurrence in order to know their real name perplexity. The perplexity classes obtained after applying the (p, γ) technique are listed in Tables 4a and 4b respectively.

perplexity class	percentage
very high (VH)	5.3%
high (H)	8.7%
medium (M)	20.9%
low (L)	27.6%
very low (VL)	37.5%

Table 4a. First Name perplexity classes

perplexity class	percentage
very high (VH)	1.8%
high (H)	3.36%
medium (M)	17.51%
low (L)	20.31%
very low (VL)	57.02%

Table 4b. Last Name perplexity classes

Tables 4a and 4b fully describe the partition for one-token names. Ordering the one token names according to their perplexity we chose the first ones according to the percentage listed above. The same process applies to the one-token last names. The values computed for two-token names are listed.

	P	γ	p_{inside}	γ_{inside}
VH	0.04%	70%	70%	80%
H	2.53%	76%	70%	80%
M	10.08%	87%	80%	82%
L	27.97%	90%	99%	90%
VL	59.38%	96.5%	99%	96.5%

Table 5. ($p, \gamma, p_{inside,} \gamma_{inside}$) values

6 CDC with Name Perplexity Estimates

The working hypothesis is that using the name partition obtained with the (p, γ) procedure we can effectively improve the accuracy of a CDC system by reducing/increasing the amount of contextual evidence required for coreferencing according to the perplexity class to each the name belong.

To construct a test corpus we have adopted the following strategy: we chose 20 two-token names such that both sets of one token-names, the first names and the last names respectively, cover the whole space in the perplexity partition. In Table 6a and 6b we present 5 first and last names used in test. As not all the 25 names formed by combining the names in 6a and 6b are found in the corpus, we consider 11 other two-tokens names having the same distribution. On the first column the names are listed, on the second column the computed perplexity (P), on the third column the number of occurrences as one-token name (O), on the fourth the number of occurrences in a two-token name (T) and on the last column the computed perplexity class (PC).

Name	P	O	T	PC
Dellai	7	31091	10722	VL
Parolari	171	1,619	2207	H
Prodi	52	9184	3382	M
Ruini	15	554	203	L
Rossi	753	7506	8356	VH

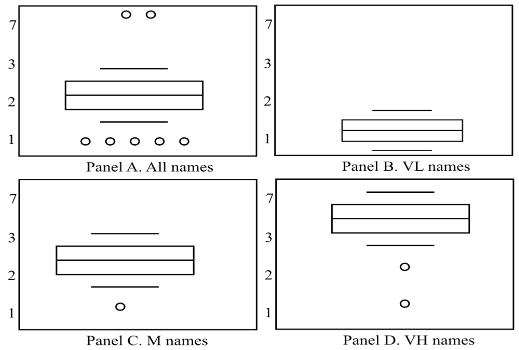
Table 6a. Test Last Names

Name	P	O	T	PC
Camillo	276	664	1731	L
Lorenzo	2088	2167	2198	H
Paolo	5255	4001	51244	VH
Romano	14	886	6414	M
Varena	5	10	85	VH

Table 6b. Test First Names

We compare the results obtained by our CDC system using the name perplexity partition (S) against two baselines: one that considers only the context at the sentence level and (BLS) one that considers the whole news (BLN). We obtain the following figures using the B-CUBED measure: S scores .72, BLS .59 and BLN .61. The gain in accuracy of more than 10% is due to the use of name perplexity classes.

The great advantage of using the (p,γ) estimates can be seen in those case where the ratio between the number of mentions and the rank of the name is close to extremes: either big number of mentions and low name perplexity, or low number of mentions and high name perplexity. In the first case the contextual evidence for coreference may be very scarce and in the second case, the requirement for strong contextual evidence is the best decision. Our results suggest that loosening the contextual requirements in the first case leads to an important gain in recall, up to 40%, while the lose in precision is less than 1.5%. The situation is best described by four panels of the five-number-summary plots of the test corpus. Panel A shows the distribution of the main five quantilies considering all the names together. Panel B shows the distribution for very low perplexity class, Panel C for medium perplexity class and Panel D for the very high perplexity class. The number of outliers in Panel A is high, which makes it difficult for any CDC system, but inside each perplexity class the variation is reduced.



Panel A. All names

Panel B. VL names

Panel C. M names

Panel D. VH names

7 Constructing an Evaluation Corpus

The (p, γ) technique could be used for constructing a test corpus for the CDC task. The main problem faced in the construction of the test corpus is data variation. The number of different entities mentioned with the same name is a random variable with a big variance. The distribution of the number of entities is very skew. The average perplexity is 2.01%, but less than 18% of the total number of names have a perplexity greater than 3. In Figure 2 we plot a modified Lorenz curve (the vertical axis is not divided in percentage, as the values are discrete).

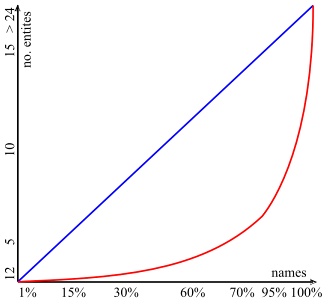


Figure 2. Lorenz Curve names/no. entities

The direct consequence of this situation is the fact that constructing an evaluation corpus by taking random names will result with a great probability in a very skew test corpus. Indeed, the expectation is that in such corpus, the average perplexity is very low, and consequently, the great majority of cases can be coreferenced by a simple algorithm. Therefore, this test corpus may be largely ineffective in ranking the algorithms. In fact, we want to construct an evaluation corpus that is able to promote the most effective algorithms. The discriminative power of a test corpus is directly related to the variance of the data. Moreover, if only certain names are considered for a test corpus, the variance can be very low; in particular, when the test corpus contains just one name the variance is zero. It is difficult to see the merits of different algorithms when tested on such corpora.

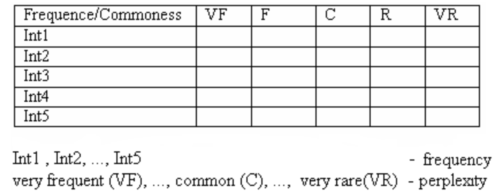
In order to make more informative statements we need to construct an evaluation corpus that is less dependent on the data variance. A possible solution is to form a partition of the set of the PNMs, that is, to split the whole set of PNMs in mutual disjunctive groups. This type of methodology is called stratified sampling, mainly because each group is a stratum. The sampling strategy, the number of sampling elements, the variance and the sampling error can be calculated independently for each strata.

The main advantages of stratified sampling are that we can concentrate on the special groups, that in general this strategy improves the accuracy of the estimation, and that the number of elements in each stratum can be conveniently chosen. The main disadvantages are related to the difficulty in finding a suitable partition of the population. The strata should be chosen prior to the sampling time, but the homogeneity inside the stratum should be guaranteed.

Our proposal is to use the name perplexity intervals. We argue that this proposal is four-fold sustainable. Firstly, the name perplexity is directly connected to the random variable whose distribution we estimate, namely the number of entities. Secondly, for free names it can be computed off - line. Thirdly, it gives us an independent and formally correct way to make a partition. Fourthly, it easily allows a separation between the important and unimportant cases.

To begin with, let us suppose we have a name that has *n* occurrences in the Adige 500K. If *n* is relatively large, than we can be sure that there are some dominant entities that may be represented by the majority of PNMs that have this name as value. However, it is unknown whether the *n* comes from the fact that there are indeed some dominant entities or whether the name by itself is a frequently used name.

In order to deal with the differences between frequency vs. perplexity, we propose to build a matrix defined by the frequency classes as rows and perplexity classes as columns. In Figure 3 we present this matrix.



Frequence/Commoness	VF	F	C	R	VR
Int1					
Int2					
Int3					
Int4					
Int5					

Int1 , Int2, ..., Int5 - frequency
very frequent (VF), ..., common (C), ..., very rare(VR) - perplexity

Figure 3. Frequency/Commonness strata matrix.

The number of different names in each of the cells of the matrix may differ according to the departure of the normal distribution of each stratum. In general, if the real distribution is normal, then as much as ten examples are sufficient. Otherwise, for not very skew distributions, which we expect most of the strata to have, an average of 30 examples should suffice. In same cases, as the normal distribution can be appropriately sampled when both *Np* and *N(1-p)* are grater than five – where p is the ratio perplexity/frequency and N the sample dimension – the number of elements in the cell may be around 200, by a maximal rough estimation.

8 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 probability belongs to the same interval. This information helps the PCDC systems lower/increase adequately the amount of contextual evidence required for coreference.

The approach presented here is effective in dealing with the problems raised by using a similarity metrics on contextual vectors improving the overall accuracy with more than 10%.

We would like to increase the number of cases considered in the sample required to delimit the perplexity classes. Equation (3) may be developed further in order to obtain exactly the number of required cases.

The (p, γ) procedure is effective is dealing with the problems regarding the construction of an evaluation corpus. The technique presented in the last section could be extended further and we are already working on a new series of experiments whose results will be made available in the near future.

Acknowledgments

The corpus used in this paper is Adige500k, a seven-year news corpus from an Italian local newspaper. The author thanks to all the people involved in the construction of Adige500k.

References

J. Artiles, Gonzalo, J., S. Sekine. 2007. *Establishing a benchmark for WePS*. In Proceedings of SemEval.

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

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

E. Elmacioglu, Y. M. F. M.Y.Khan, D. Lee. 2007. *PSNUS: Web People Name Disambiguation by Simple Clustering with Rich Features,* in Proceedings of SemEval

H. Han, W. Xu. 2005. *A Hierarchical Bayes Mixture Model for Name Disambiguation in Author Citations*, in Proceedings of SAC'05

R. Hog, J. McKean, A. Craig, 2006. *Introduction of Mathematical Statistics*, ed. Prentice Hall

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

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

Q. Vu, T. Massada, A. Takasu, J. Adachi. 2007. *Using Knowledge Base to Disambiguate Personal names in Web Search Results*, In Proceedings of SAC