Dynamic Parameters for Cross Document Coreferece

Octavian Popescu
papsi@racai.ro
Racai, Romanina Academy

Abstract

In this paper we present a new algorithm for the Person Cross Document Coreference task. We show that accurate results require a way to adapt the parameters of the similarity function – metrics and threshold – to the ontological constraints obeyed by individuals. The technique we propose dynamically changes the initial weights computed when the context is analyzed. The weight recomputation is necessary in order to resolve clusters borders, which are inevitably blurred by a static approach. The results show a significant gain in accuracy.

1 Introduction

The Person Cross Document Coreference, CDC, task requires that all the personal name mentions, PNMs, in a corpus be clustered together according to the individuals they refer to (Grishman 1994). The coreference between two PNMs is decided on the basis of the local contexts. In this paper we consider a news corpus, and the local context is the piece of news to which a particular PNM belongs. We work on a seven year Italian local newspaper corpus, Adige 500K (Magnini et. al. 2006).

While there are certain similarities between a disambiguation task and the CDC task, we maintain that there is a significant difference which sets the CDC task apart. 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 allows for disambiguation decisions which are made by considering the local context only. On the other hand, the distribution of the PNMs in a corpus is rather random and the relevant coreference context is a dynamic variable which depends on the diversity of the corpus, that is, on how many different persons with the same name share a similar context. Unlike the word senses which are subject to strong linguistic constraints, the name distribution is more or less random. 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 paper we present a method to exactly determine the relevance of a piece of context for the coreference. As above, the exactness is understood in relationship with the whole system of clusters. The relevance of a piece of context is computed by means of a weighting procedure. The classic weighting procedures are static, each piece of context receives an initial value that is also a final one and the clustering proceeds on the basis of these values. We demonstrate that this approach has serious drawbacks and we argue that in order to obtain accurate results, a dynamic weighting procedure is necessary, which outputs new values depending on the cluster configuration.

In Section 2 we review the relevant literature. In Section 3 we present the problems related to the classical approach to the CDC task and we present evidence that the data distribution in a news corpus requires a proper treatment of these

Coling 2008: Poster Volume, pages 988–996,
Beijing, August 2010

problems. In Section 4 we present the technique that permits to overcome the problems identified in Section 3. In Section 5 we present the context extraction technique that supports the method developed in Section 4. In Section 6 we present the results of an evaluation experiment. The paper ends with Conclusion and Further Work 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, by 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 Smiths" 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 very often found 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". These types of name have different cluster systems

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 CDC 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 that have 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 big the threshold should be. The system of clusters is not developed unrealistically if we are able to handle the fact that individuals obey different constraints which are derived directly from the ontological properties. These constraints are determined directly from the context and adequate weights can be set.

Recently, there has been a major interest in the CDC systems, and, in the last two years, two important evaluation campaigns have been organized: Web People Search-1 (Artiles et al. 2007) and 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 3.2). For reliable results, a CDC system must have access to global information regarding the coreference space.

Rich biographic facts have been shown to improve the accuracy of CDC (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 CDC systems based on VSM, 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).

3 Coreference Based on Association Sets

The coreference of two PNMs is realized on the basis of the context. In a news corpus, the context surrounding each PNM, which is relevant for coreference, is extracted into a set, called association set. In Table 1 we present an example of association sets related to the same name.

Name	Associated Sets
Paolo Rossi	TV, comedian, , satire research, conference politics, meeting

Table 1: Associated Sets

A weighting schema, a global metrics and threshold are set, and the distance between two association sets is computed. The decision of coreferencing two PNMs is made on comparing the distance to the threshold and clustering the PNMs representing the same individual into a unique cluster. The accuracy of a CDC system based on association sets depends on two factors: (1) the ability to extract the relevant elements for the association sets from the news context and (2) the adequacy of the similarity formula - metrics and threshold.

Regarding the first factor, the ability to extract the relevant pieces of context, the right heuristics must be found, because the exact syntax-semantics analysis of text is unfortunately very hard or impossible to implement. A strong limitation comes from the fact that even a shallow parsing requires too much time in order to be practical. However, it has been shown that accurate parings of PNMs and co-occurring special words can be found by employing relaxed extraction techniques (Buitelaar&Magnini 2005). The association sets built in this way are effective in solving the CDC task (Sekine 2008, Popescu 2008). We make use of these findings in order to build the association sets, which mainly include named entities and certain special words, which are bound to an ontology. The details of these particular association sets are given in Section 5.

As straightforward as the classical approach based on the distance between association sets may seem, there are actually a series of problems related to the second requirement, namely the adequacy of similarity formula. We make these problems explicit below.

3.1 Masking, Superposition and Border Proximity

In order to introduce the first problem we start with an intuitive example. Suppose that we want to individuate the persons with the name Michael Jackson in a news corpus. A simplistic solution is to cluster together all such PNMs and declare that than there is just one person mentioned in the whole corpus with this name. This solution has the advantage of being very simple and of obtaining a very high score in terms of precision and recall. This is because most of such PNMs refer to only one person indeed – the pop star. However, the above method fails short when it comes to presenting the evidence for its coreference decision. Actually, it turns out that this is a very hard task, because the number of PNMs, which do not refer to the pop star, is extremely small. Thus, the prior chances of correctly finding two PNMs which do not refer to this person are quite small. Unfortunately, the classical metrics are too coarse to capture the difference in such cases, even if the association sets are 100% correct. To support this statement, let us consider three classes under the same name, with each class corresponding to a different individual. Let us further suppose that two classes contain the great majority of the PNMs, and the third class only has a small number of PNMs. A linear decision is likely to confound the elements of the third class to the ones of the first two[1]. This happens because the elements of the third class are transparent to the hyper plane that separates the two well-represented classes. This situation is called masking, and is a direct effect of applying an inaccurate weighting schema and metrics (Hastie&Tibshirani 2001). The effects of masking on the CDC task have been empirically noticed in (Pederson 2005). The main obstacle in dealing with masking is the correct treatment of the border elements. δ_{ij}, the discrimant function between two classes, i and j respectively, must assign zero to all border elements. In Section 4, we directly address this problem.

The second problem that needs to be solved by the CDC systems based on associated sets may be regarded as the negative effect of counter balancing the sparseness problem. In general, the association sets are too sparse to permit pair to pair comparison. Rather, the information must be interpolated from a set of corefered association sets. For example, in Figure 1, any two association sets chosen from the three ones on the left, AS_1, AS_2 and AS_3 respectively, are similar

[1] In fact any decision functions that can be bijectively transformed into a linear function, like most exponential kernel functions for example, are similarly prone to masking.

enough to one another to corefer. However, none of these association sets is similar enough to the one on the right – AS_4. But accepting the coreference of any initial pair, in this particular case, we implicitly accept the coreference with the fourth one.

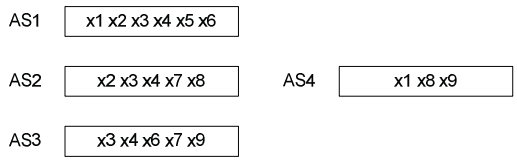

Figure 1. Interpolating

By interpolating the information in the set of the initial three association sets, the coreference becomes possible between all four association sets. In general, by interpolating from a set of the association sets, one wants to find the right coreferences and to avoid the false ones accurately. In a vector space, the interpolation is safe if the initial vectors are orthogonal to each other, because the sum of orthogonal vectors is also orthogonal to any other vector that is not part of the sum. Therefore the right coreferences have a big dot product with the sum, while the false ones have a dot product with the sum close to zero. This property of the sum of the orthogonal vectors is called superposition (Gallant 1993). By representing the association sets as vectors, where each set of vectors is associated exclusively with a certain individual, the sum of these vectors has the superposition property.

However, if the vectors representing the association sets are not orthogonal, then the interpolated vectors are prone to false coreferences. In this case, the accidental coincidences – which are responsible for the original vectors not being orthogonal – biases the dot product and introduces false coreferences. Consequently the superposition affects negatively the overall accuracy. The aggravating effect of superposition in conjunction with an agglomerative clustering procedure has been empirically noted in Gooi&Allan.

The third problem is directly related to the fact that in the most ambiguous cases the association sets lead to high-dimensional, very sparse vectors. The basic fact is that inside a cluster of correctly corefered PNMs that refer to the same individual, the distance from most of these PNMS to the center of the cluster is smaller than the distance from these PNMs to the border. Let us consider that all the *m* PNMs representing the same individual are points in an *n* dimensional vector space and their cluster is normalized to the unit sphere. The distance from the center of the sphere to the closest point is an exponentially growing formula both in *1/n* and *1/m*. Even for small values, the distance from the center to the closest point is larger than ½. The points representing the PNMs in the same cluster are closer to the border, and not to the center of the sphere. This is a secondary effect of the curse of dimensionality problem in the vector space[2].

3.2 Data Distribution

Let us consider the corpus, focusing on the distribution of PNMs. Many PNMs are the mentions of the same name, considered as a string. We are interested in the frequency with which a certain name appears. We have noticed that there is a strict relationship between the names, their frequencies and the number of mentions; see Table 2.

Freq	PNM	# PNM
1	317,245	317,245
2 – 5	166,029	467,560
6 – 20	61,570	634,309
21 – 100	25,651	1,090,836
101 – 1000	7,750	2,053,994
1001 – 2000	4,25	569,627
2001 – 4000	157	422,585
4001 – 5000	17	73,860
5001 – 31091	22	190,373

Table 2 Frequency of Names and PNMs in Adige500k

The names have a very unbalanced distribution. A name which has a frequency over 20 and is ambiguous represents a difficult case. The measure we use in order to evaluate the difficulty is the Gini's mean difference. Let X_1, X_2, …, X_n be the individuals that are named with the same name and let S be the set of the PNMs of this name PNMS, S_1, S_2, … S_n. The Gini's mean difference is a measure of the spread of the information in the set S:

$$\sum_{j=2}^{n} \sum_{i} \binom{n}{2} |Si - Sj| = G \qquad (1)$$

The uniform distribution makes Gini's factor null. A value of this factor close to 1 shows a skewed distribution. In the first case, $G \approx 0$, the superposition effect is likely to be responsible for false coreferences, while in the latter case, $G \approx 1$,

[2] The curse of dimensionality refers to the fact that the number of sample points required to state confident values for a statistics grows exponentially with the dimension of vector space.

the masking effect is predominant. However, there is a close relationship between all the three problems above. As the most ambiguous cases are near the border, it is likely that the vectors are not orthogonal and consequently the false coreferences are introduced in the system, which ultimately leads to masking.

4 Resolving the Border Condition

We are going to present a technique developed to deal with the problems identified in the previous section. The bottom line is that the weights and the threshold required by the similarity function of two association sets should be dynamically computed. In this way the border between any pair of clusters can be accurately set.

We present the procedure of adjusting the weights and the threshold for a given group of clusters in order to maximize the probability of the correct coreferences. The first step is to present the construction of the association sets, with initial weight values. The second step is to show how these initial weight values are recomputed for a set of given clusters.

Initialization

As mentioned in the first paragraph of Section 3, the association sets are built out of the surrounding context by considering the named entities, and special words. The named entities are clearly marked in the input, the corpus having being tagged by a Named Entities Recognition tool. The words considered special are identified using an ontology and the procedure is given in Section 5. The construction of the association set is a search procedure starting from the PNM. The first search space is the longest nominal group which is headed by a PNM:

uno dei falchi dell' amministrazione di Stati Uniti guidata dal presidente George W.Bush
one of the falcons of the U.S. administration lead by the president Georg W. Bush

All the special words that are present in this nominal group are included in the association set of this PNM. In this example, these special words are "president" and "administration" respectively. The named entity "U.S." is also included. These elements receive the highest weights. The search space is extended to the sentence level and new named entities/special words are included. However, unlike in the first phase, the weight of these words is determined on the basis of a second parameter, namely the number of different names interfering between the PNM and these words. We take into consideration three values 0, 1 and 2 or more. After the sentence, the next search domain is the whole news. Basically, the significance of an element decreases linearly with the distance and the number of other interfering PNMs. In Table 3 we present the linear kernel weighting schema described above. The series α_{ij} is decreasing linearly over both indexes.

	Interfering PNMs		
Domain	0	1	≥2
PNM Group	α_{11}	α_{12}	α_{13}
IN Sentence	α_{21}	α_{22}	α_{23}
Out Sentence	α_{31}	α_{32}	α_{33}

Table 3. Linear Kernel for Initial Weights

Recomputation

The association set is basically a pair of two vectors: $X = (x_1, \ldots, x_n)$ the set of words and $W = (w_1, \ldots, w_n)$ the set of the initial weights. Two PNMs corefer or not depending on whether the sum of their common part is bigger, respectively lesser than a threshold.

$$coref: \sum_{common\ xi} wi \geq T \quad (2)$$

$$non\ coref: \sum_{common\ xi} wi \leq T \quad (3)$$

Suppose now that we have an independent way to know the truth regarding the coreference. Then, we have to readjust the initial weights such that the real configuration of clusters is promoted also by Equations (2) and (3). For clarity, let us give an example: suppose that we know that in our corpus there is only one person named "Roberto Bizzo" and only one person named "Roberto Cuillo", and no other person is called "Roberto". Consequently the PNMs "Roberto" are clustered to the clusters "Robert Bizzo" xor "Roberto Cuillo". Suppose further that the named entity "Roma" is associated with some of the PNMs "Roberto". If only "Roberto Bizzo" is associated with "Roma", then the coreference between those "Roberto" associated with "Roma" and "Roberto Bizzo" can be made. However, it is often the

case that both "Roberto Bizzo" and "Roberto Cuillo" are associated with "Roma", which has its particular weight for each PNM. In this case this named entity, "Roma", may bear no relevance for the coreference of "Roberto" in either of the clusters. Consequently, whatever the initial value for "Roma" in certain association sets, it must be nullified. In order to find out which elements of the association sets are relevant, and what weights the relevant elements must have, we propose the following strategy: we replace the "Roberto Bizzo" with "Roberto X", and "Roberto Cuillo" with "Roberto X". We obtain a big set of association sets corresponding to the PNMs "Roberto X". We reweight the elements of their association sets and the threshold, such that, from this set of association sets, we obtain exactly two clusters, one that is identical with "Roberto Bizzo", and one that is identical with "Roberto Cuillo". Conceptually, this strategy is similar to the pseudo words technique used in building test corpora. After the reweighting of the elements associated with "Robert Bizzo" and "Roberto Cuillo" respectively, we can associate the simple PNMs "Roberto" to one of these two clusters.

In the above example we make use of the fact that if two persons have different last names then they are different persons. This is a prior ontological constraint. In fact, whenever we know the set of ontological constraints that correctly cluster a set of PNMs in two or more clusters, we can intentionally confound the PNMs, recompute the weights and the thresholds of their association sets, in order to obtain the initial cluster configuration. Now we use the new computed values to cluster new PNMs whose relationship with the ontological constraints could not have been determined from the corpus.

We show that we can use the Simplex method to recompute the initial weights. Indeed, by intentionally confounding a system of clusters, we determine the coefficients which, when multiplied with the initial weights, lead to the correct clustering. These coefficients are the solution to a set of inequalities like those presented in Equations (2), and (3). The objective function in Simplex is a max or a min depending on whether we know that the PNMs corefer or not: if they do not corefer then there is a max Simplex system, and the threshold is just higher than the value of the objective function. Let us give an example. Suppose we have the following configuration, where AS_i represents the association set of the PNM_i, where wi is the vector of the initial weights and T is the threshold: .

$AS_1 \cap AS_2 = \{x_1, x_2, x_3\}$	wi = (1, 2, 2)	T = 7	No Coreference	$x_1 + 2x_2 + 2x_3 \leq 7$
$AS_1 \cap AS_3 = \{x_1, x_3\}$	wi = (5,0,4)	T = 11	Coreference	$5x_1 + 4x_3 \geq 11$
$AS_2 \cap AS_4 = \{x_2, x_3\}$	wi = (0,3,4)	T = 9	Coreference	$3x_2 + 4x_3 \geq 9$
$AS_5 \cap AS_6 = \{x_1, x_2\}$	w i= (2,1,0)	T = ?	No Coreference	max $(2x_1 + x_2)$

The above cluster configuration leads to the following Simplex system:

$$\max 2x_1 + x_2$$

$$\begin{cases} x1 + 2x2 + 2x3 \leq 7 \\ 5x1 + 4x3 \geq 11 \\ 3x2 + 4x3 \geq 9 \end{cases}$$

which has the solution wr = (1.55, 1.91, 0.82) with max = 5. Therefore the initial weights for the elements x_1, x_2, x_3 must be multiplied with 1.55, 1.91, 0.82 respectively and the appropriate threshold for making a decision is 5.01.

5 Ontological Constrained Association Sets

In the preceding section we presented a strategy based on Simplex Algorithm developed for the border weight assignment. The similarity formula is recomputed such that a set of ontological restriction is satisfied. In this section we present the way the set of ontological restrictions is found. The set of special words is identified on the basis of an ontology. We have used SUMO (Niles 2003) because it has the advantage that its hierarchies are connected to the WordNet, which is a Multilanguage aligned resource. Below we present the main categories of the SUMO attributes used. Summing up, there are more than 7 000 special words taken into account.

Corporation
Organization
Occupational Role
Occupies Position
Social Interaction
Social Role
Unemployed

There are mainly three different ways to create the set of ontological restrictions: fixed, prior ontological constraints, local restrictions and exclusive ontological relationships.

The fixed, prior ontological constraints are those that tend to be expressed in a fixed pattern, making it easy to identify them in the context. Usually they express the date and place of birth,

contact information, but also the gender, the family relationship, the ethnic group etc.

The local restrictions are a very rich source of information. It has been argued that inside each piece of news the coreference of all the PNMs is a valid procedure, with more than 99% accuracy (Popescu et al. 2008). By comparing the structure of the largest nominal group headed by two locally corefered PNMs we can found ontological compatibilities. Table 4 shows a sample of the compatible pairs as extracted from corpus. These pairs can be used successfully for coreferencing purposes, but these do not form ontological hierarchies and cannot be used to build inference chains.

Pairs of compatible professions
albergatore comerciante
ala giocatore
agronomo professore
allenatore mister
alpinista guida alpina
architetto progettista
arcivescovo monsignore
monsignore teologo
monsignore sacerdote
assessore consigliere

Table 4. Compatible Occupational Role

The exclusive ontological relationships are given explicitly under the form of rules. These rules stipulate what is ontologically inacceptable. We have seen an example of such rules referring to the family names in Section 4. The Occupational Role and Social Role attributes are one of the most useful exclusive ontological ones, because they are frequently mentioned in a news corpus. In average, local information at the news level produces a special word from the above categories in approximately in 30% of cases (Magnini et al 2006.). An example of the realization of the exclusive rules for a sample of multi pairs of words as extracted from corpus is presented below:

Secretary≠Priest≠Judge
Architect≠Attorney
Waiter≠Manager
Actor≠Researcher

The system of clusters determined using the technique described in Section 4 obeys the set of these constraints. The set C of ontological constraints are used to generate active rules at the word level, which, by means of fixed text patterns, are compared against the association sets. This permits the realization of ontological motivated cluster systems, which in combination with the technique of reweighting presented, leads to accurate new coreferences outside the scope of C, while avoiding the border problems presented in Section 3 ..

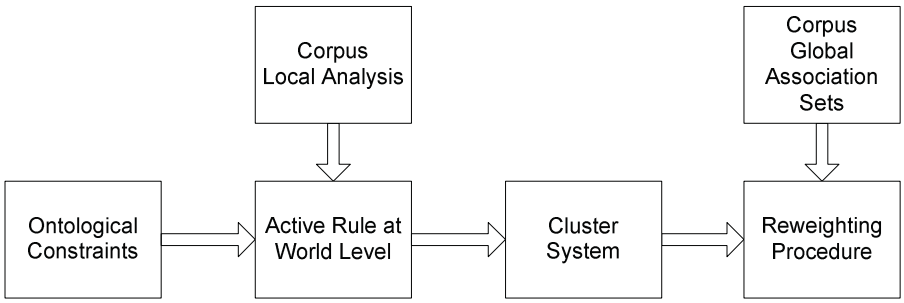

Figure 2. The dynamic reweighting schema flow

6 Evaluation

The technique we propose is designed for an accurate border detection between clusters of ambiguous names. We created a sample of the ambiguous names. For each name we computed the Gini's mean difference using the formula introduced in Section 3, which gives an indication of the spread of information relevant for coreference. We have noticed that there is a strong correlation between the Gini's mean difference and the difficulty of a coreference system. The names chosen for this experiment are such that the Gini's factor uniformly distributed in (0,1). However, the number of PNMs for each name is bigger than the number of individuals having that name. The choice is motivated by the fact that these are the most difficult cases for a CDC system, as they require strong and consistent evidence for accurate results. The opposite cases, when the number of the individuals is close to the number of PNMs or the Gini's coefficient is

close to 0 or 1, can be approached with a pure statistical approach (Popescu 2009).
The first column in Table 5 lists the names, the second column lists the number of the PNMs considered for each name, the third column lists the number of individuals having the respective name, the fourth column lists the number of PMNs for each individual, the fifth column lists the Gini's factor and the sixth column lists how many clusters have been found obeying ontological constraints/ and how many PNMs have been clustered in these clusters.

Name	#PNMs	#P	Distribution	Gini	Constraints
Angelo Elia	58	5	{20,24,7,2,2,3}	.428	2 / 18
Gifuni	89	3	{47,21,31}	.175	3/ 12
Giuseppe Rossi	185	12	{69,32,5,9,4,5,6,6,12,7,8,22}	.503	5 / 38
Paulo Rossi	137	9	{91,17,9,3,2,3,5,5,2}	.673	3 / 74
Schlesinger	62	4	{26,19,6,11}	.274	4 / 19
Tanzi	370	3	{315,49,16}	.524	3/129

Table 5. Name Test Set

We compare the technique proposed in Section 4 (DYN) against three different approaches: the first is a no weight coreference, requiring a fix number of similar elements in the association set (NOW), the second is Baga&Baldwin quadratic metric formula at sentence level (BB), and the third is an agglomerative vector space clustering algorithm as in Gooi&Allan(GA). All these three approaches use fixed similarity parameters.
The evaluation is done using the B-CUBED algorithm (Baga&Baldwin). The results, computed with F formula, are presented in Table 6.

Name	NW	BB	GA	DYN
Angelo Elia	.426	.639	.684	.672
Gifuni	.53	.635	.661	.726
Giuseppe Rossi	.481	.619	.589	.673
Paulo Rossi	.446	.623	.598	.691
Schlesinger	.528	.588	.723	.829
Tanzi	.572	.539	.699	.815
Average	**.417**	**.607**	**.659**	**.734**

Table 6. F-formula on B-CUBED

The BB and GA have been tested on the John Smith corpus, which contains the PNMs of just one name, John Smith. As John Smith is a very common name and no famous person carries it, this corpus is rather biased as the Gini's factor is small; that is why BB performs better than GA on "Giuseppe Rossi" and "Paulo Rossi". The DYN scores the best , gaining in average 7 points in F formula.

Conclusion and Further Work

In this paper we present a new technique for the CDC task which allows us to dynamically change the weights in the association sets in order to accurately account for border cases. As we showed in Section 3, the border cases are actually the most important ones due to the high dimensionality of the vector space which models the association sets.

The results we have obtained are superior to other approaches. We think that this is possible because the technique we used directly addresses the problem related to masking and superposition.

We plan to further study this technique by following mainly three directions. First, we want to study further the behavior of masking and superposition within a larger test corpus. Second, we want to extend the set of exclusive ontological relationships which can be determined from the context with shallow text analysis. Third, we want to understand better the ways in which the set of ontological constraints interact with the vector space in order to increase the overall accuracy of the coreference system.

A secondary effect of the proposed technique is that a stronger control of the inferences resulting from a cluster system can be obtained. In the future this seems to be a promising method to link the coreference tasks to the chain of inferences.

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.

G. Mann, D. Yarowsky. 2003. *Unsupervised Name Disambiguation*, in Proceeding of HLT-NAACL

I. Niles, A. Pease, 2003. *Linking Lexicons and Ontologies: Mapping WordNet to the Suggested Upper Merged Ontology*, in Proceeding IKE

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

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 2009. *Name Perplexity*. In Proceedings of NAACL HLT

P. Buitelaar, B. Magnini (Eds.) 2005. *Ontology Learning from Text: Methods*, Evaluation and applications. *IOS Press*

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

T. Hastie, R. Tibshirani, J. Friedman, *2001. The elements of Statistical Learnig*, Springer Press

S. Gallant, *Neural Network Learning*, MIT Press

S. Sekine, 2008 *Extended Named Entity Ontology with Attribute Information*, in Proceeding of LREC

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