IRST-BP: Web People Search Using Name Entities

Octavian Popescu
FBK-irst, Trento (Italy)
popescu@itc.it

Bernardo Magnini
FBK-irst, Trento (Italy)
magnini@itc.it

Abstract

In this paper we describe a person clustering system for web pages and report the results we have obtained on the test set of the Semeval 2007 Web Person Search task. Deciding which particular person a name refers to within a text document depends mainly on the capacity to extract the relevant information out of texts when it is present. We consider "relevant" here to stand primarily for two properties: (1) uniqueness and (2) appropriateness. In order to address both (1) and (2) our method gives primary importance to Name Entities (NEs), defined according to the ACE specifications. The common nouns not referring to entities are considered further as coreference clues only if they are found within already coreferred documents.

1 Introduction

Names are ambiguous items (Artiles, Gonzalo and Sekine 2007). As reported on an experiment carried out on an Italian news corpus (Magnini et all 2006) within a 4 consecutive days from a local newspaper the perplexity is 56% and 14% for first and last name respectively. Deciding which particular person a name refers to within a text document depends mainly on the capacity to extract the relevant information out of texts when it is present[1]. We consider "relevant" here to stand primarily for two properties: (1) uniqueness and (2) appropriateness. A feature is unique as long as it appears only with one person. Consider a cluster of web pages that characterizes only one person. Many of the N-grams in this cluster are unique compared to other cluster. Yet the uniqueness may come simply from the sparseness. Appropriateness is the property of an N-gram to characterize that person.

Uniqueness may be assured by ontological properties (for example, "There is a unique president of a republic at a definite moment of time", "Alberta University is in Canada). However, the range of ontological information we are able to handle is quite restricted and we are not able to realize the coreference solely relying on them. Uniqueness may be assured by estimating a very unlike probability of the occurrence of certain N-grams for different persons (as, for example, "Dekang Lin professor Alberta Canada Google").

Appropriateness is a difficult issue because of two reasons: (a) it is a dynamic feature (b) it is hard to be localized and extracted from text. The greatest help comes from the name of the page, when it happens to be a suggestive name such as "homepage", "CV", "resume" or "about". Gene-

[1] It is very difficult to evaluate whether the information allowing the coreference of two instances of a (same) name is present in a web page or news. A crude estimation on our news corpus for the names occurring between 6-20 times, which represent 8% of the names inventory for the whole collection, is that in much more than 50% of the news, the relevant information is not present.

Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007), pages 195–198,
Prague, June 2007.

alogy pages are very useful, to the extent that the information could be accurately extracted and that the same information occurs in some other pages as well. However, in general, for plain web pages, we rely on paragraphs in which a single person is mentioned and consequently, the search space for similarity is also within this type of paragraphs.

Our proposal is to rely on special N-grams for coreference and it is a variant of agglomerative clustering based on social networks(Bagga&Baldwin 1998, Malin 2005) . The terms the N-grams contain are crucial. Suppose we have the same name shared by two different persons who happen to also have the same profession, let's say, "lawyer", and who also practice in the same state. While all three words – (name, profession, state) - might be rare words for the whole corpus, their probability computed as chance to be seen in the same document is low, their three-gram fails to cluster correctly the documents referring to the two persons[2]. Knowing that the "lawyer" is a profession that has different specializations, which are likely to be found as determiners, we may address this problem more accurately considering the same three-gram by changing "lawyer" with a word more specific denoting her specialization.

The present method for clustering people web pages containing names according addresses both uniqueness and appropiateness. We rely on a procedure that firstly identifies the surest cases of coreference and then recursively discover new cases. It is not necessarily the case that the latest found coreferences are more doubtful, but rather that the evidence required for their coreference is harder to achieve.

The cluster metrics gives a primary importance to words denoting entities which are defined according to ACE definitions: PER, LOC, ORG, GPE.

In Section 2 we present in detail the architecture of our system and in Section 3 we present its behavior and the results we obtained on the test set of Semeval 2007 Web Person Search task. In section 4 we present our conclusions and future directions for improvement.

[2] The traditional idf methods used in document clustering must be further refined in order to be effective in person coreference.

2 System Architecture

First, the text is split into paragraphs, based mainly on the html structure of the page. We have a Perl script which decides weather the name of interest is present within a paragraph. If the test is positive the paragraph is marked as a person-paragraph, and our initial assumption is that each person-paragraph refers to a different person.

The second step is considered the first procedure of the feature extraction module. To each paragraph person we associate a set of NEs, rare words and temporal expressions, each of them counting as independent items. For all of these items which are inside of the same dependency path we also consider the N-grams made out of the respective items preserving the order. For each person-paragraph we compute the list of above items and consider them as features for clustering. This set is called the association set.

The first step in making the coreference is the most important one and consists in two operations: (1) the most similar pages are clustered together and (2) for each cluster, we make a list of the pages which most likely do not refer to the same person. Starting with this initial estimation, the next steps are repeated till no new coreference is made.

For each cluster of pages, a new set of items is computed starting from the association sets. Only the ones which are specific to the respective cluster - comparing against all other clusters and against the list of pages not related (see (2) above) – are kept in the new association set. These are the features we use further for clustering. The clustering score of two person-paragraphs is given by summing up the individual score of common features in their association sets. The score of a feature is determined based on its type - (NE, distinctive words, temporal expressions) - , its length in terms of words compounding it, and the number of its occurrences inside the cluster and inside the whole corpus, considering only the web pages relative to that name and the absolute frequency of the words. The feature score is finally weighed with a factor which expresses the distance between the name and the respective feature. An empirical threshold has been chosen.

Each of the above paragraphs representing a module in our system is explained in one of the next subsections respectively.

2.1 Preprocessing

Web pages contain a lot of information outside the raw text. We wrote Perl scripts for identifying the e-mail addresses, phone and fax numbers and extract them if they were in the same paragraph with the name of interest. It seems that a lot can be gained considering the web addresses, the type of page, the links outside the pages and so on. However, we have not exploited up to now these extra clues for coreference. The whole corpus associated with a name is searched only once. If the respective items are found in two different pages, these two pages are clustered.

In web pages, the visual structure plays an important role, and many times the graphics design substitutes for linguistics features. Using a normal html parser, such as lynx, the text may lack its usual grammatical structure which may drastically decrease the performances of sentence splitters, Name Entity Recognizers and parsers. To alleviate this problem, the text is first tagged with PoS. If a paragraph, '\n', does not have a main verb, then it is treated separately. If the text contains only nouns and determiners and if the paragraph is within a paragraph containing the name of interest, the phrase "You are talking about" is added in front of it to make it a normal sentence.

The text is split into person-paragraphs, and each person-paragraph is split into sentences, lemmatized, the NEs are recognized[3] and the text is parsed using MiniPar (Dekang Lin 1998). We are interested only in dependency paths that are rooted in NEs – the NP which are included in bigger XP, or sister of NPs, or contain time expressions.

The person-paragraphs are checked for the interest names. We write rules for recognizing the valid names. If a page does not have a valid name of interest, it is discarded. A page is also discarded when a valid name of interest has its entity type "ORG".

[3] We thank to the Textec group at IRST for making it possible for everyone to pre process the text very easily with state of the art performances.

2.2 Feature Extraction

The association set contains a set of features. The features are NEs or part of NEs, because the closed class words, the very frequent words – computed on the set of all web pages for all persons – are deleted from the NEs[4]. When we refer to the length of a feature we mean the number of words it is made of, after deletion.

We consider words (phrases) which are not NEs as features but only if they are frequent in already coreferred person-paragraphs. That is, initially the coreference is determined solely on NEs. If there is enough evidence, i.e. when a word is frequent within the cluster and not present within other clusters, then the respective word (phrase) is taken into account for coreference.

Time expressions are relevant indicators for coreference if they are appropriately linked to a person. We consider them always, just like a NE, but when they appear in particular dependency trees they have a special value. If they are dominated by a name of interest and/or by the lemma "birth", "born" we consider them as a sure factor for coreference.

For all composed features we also consider the order preserved combinations of their parts obtaining new features.

The association sets increase their cardinality by coreference. At each step, the new added features are checked against the ones from the other clusters. The common features are kept in separate sets. The coreference is not decided on their basis, but these features are used to identify the paragraph persons that do not refer to a particular person, and therefore should not be included in the same cluster. We do not explicitly weigh differently the features (apart of the cases mentioned above) but they are actually weighed differently implicitly. The words within a composed feature are repeated, a feature of length n produces n(n-1) new features, n> 2. Besides, as we will see in the next section, the similarity score uses the length of a feature.

[4] Sometimes, correctly or not, the SVM base NER we use includes, especially inside of LOC and GPE name entities, common words. In order to remain as precise as possible, we choose not to consider these words when we compute the similarity score.

2.3 Similarity Measure

Our similarity score for two person-paragraphs is the sum of the individual scores of the common features which are weighed according to the maximum of distances between the name of interest and the feature.

There are three parameters on which we rely for computing similarity: the length, the number of occurrences, and the absolute frequency of a feature. The score considers the cube of the feature length (which means that the one word features do not score). We compute the ratio between the number of occurrences within the cluster and the number of occurrences in the web pages relative to that name. The third parameter is the absolute frequency of the words. As usually, if the word is a rare word it counts as more evidence for coreference. We regard these parameters as independent, in spite of their relative dependency, and we simply multiply them.

We define the distance between a feature and a name as a discrete measure. If the name and the feature are sisters of the same head then their distance is minimum, therefore their importance for similarity is the highest. The second lower distance value is given within the same sentence and the distance increases with the number of sentences. If there are no other names mentioned in the paragraph, the distance is divided by half.

We have established an empirical threshold which initially is very high, as the features are not checked among the clusters in the first run. After the first run, it is relaxed and the common and individual sets are computed as we have described in the previous section.

3 Evaluation

The system performance on the test set of Semeval 2007 Web Person Search task is $F_{\alpha=0.5}$ = 0.75, harmonic means of purity, and $F_{=0.2}$ = 0.80 - the inverse purity mean. The data set has been divided in three sets: SET1 ACL people, SET2 Wikipedia people, and SET3 census people. The results are presented in table 1. The fact that the system is less accurate on SET2 may be due to the fact that larger person paragraph are considered and therefore more inappropriate similarity are declared.

Test Set	Purity	Inverse Purity	$F_{\alpha=0.5}$
SET1	0,75	0,80	0,77
SET2	0,83	0,71	0,77
SET3	0,81	0,75	0,78

4 Conclusion and Further Research

Our method is greedy and it depends a lot on the accuracy of coreference as the system propagates the errors from step to step.

One of the big problems of our system is the preprocessing step and further improvement is required. That is because we rely on the performances of NER and parsers. We also hope that by the inclusion of extra textual information the html carries, we will have better results.

A second direction for us is to exactly understand the role of ontological information. For the moment, we recognized some of the words denoting professions and we tried to guess their determinators. We think that having hierarchical relationships among LOC, GPE and also for ORG may make a difference in results especially for massive corpora.

References

Artiles, J., Gonzalo, J. and Sekine, S. (2007). *Establishing a benchmark for the Web People Search Task: The Semeval 2007 WePS Track.* In Proceedings of Semeval 2007, Association for Computational Linguistics.

Bagga A., Baldwin B.,(1998) *Entity-Based cross-document-referencing using vector space model,* In proceedings of 17th International Conference on Computational Linguistics

Magnini B., Pianta E., Popescu O. and Speranza M. (2006). Ontology Population from Textual Mentions: *Task Definition and Benchmark. Proceedings of the OLP2 workshop on Ontology Population and Learning, Sidney, Australia,*. Joint with ACL/Coling

Malin. B., (2005): *Unsupervised Name Disambiguation via Network Similarity*, In proceedings SIAM Conference on Data Mining 2005

Zanolli R., Pianta E. (2006) *Technical report*, ITC IRST