Semi-automatic labeling of coreferent named entities: an experimental study

Erwan Moreau, François Yvon, Olivier Cappé

▶ **To cite this version:**

Erwan Moreau, François Yvon, Olivier Cappé. Semi-automatic labeling of coreferent named entities: an experimental study. LREC 2008 workshop W15: 'Resources and Evaluation for Identity Matching, Entity Resolution and Entity Management', 2008, Marrakech, Morocco. pp.1-7. hal-00487082

HAL Id: hal-00487082
https://hal.science/hal-00487082

Submitted on 27 May 2010

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Semi-automatic labeling of coreferent named entities: an experimental study

Erwan Moreau, François Yvon, Olivier Cappé

Télécom ParisTech & LTCI/CNRS, Univ. Paris Sud & LIMSI/CNRS, Télécom ParisTech & LTCI/CNRS
erwan.moreau@enst.fr, yvon@limsi.fr, cappe@enst.fr

Abstract

In this paper, we investigate the problem of matching coreferent named entities extracted from text collections in a robust way: our long-term goal is to build similarity methods without (or with the minimum amount of) prior knowledge. In this framework, string similarity measures are the main tool at our disposal. Here we focus on the problem of evaluating such a task, especially in finding a methodology to label the data in a semi-automatic way.

1. Introduction

In this paper, we study the problem of matching coreferent named entities in text collections, focusing primarily on orthographical variations in nominal groups (i.e. we do not handle the case of pronominal references). As described in the literature (e.g. (Christen, 2006)), textual differences between entities are due to various reasons: typographical errors, names written in different ways (with/without first name, with/without title, etc.), abbreviations, lack of precision in organization names, etc. Among them, we are particularly interested on capturing textual variations that are due to transliterations (translations between different alphabets). Identifying textual variations in entities is useful in many text mining and/or information retrieval tasks. In the former case, it will act as a useful normalization step, thus limiting the growth of the indexing vocabulary (see e.g. (Steinberger et al., 2006)). In the latter case, for instance, it allows to retrieve relevant documents even in the face of misspelling (in the query or in the document).

There are different ways to tackle the problem of NE matching: the first and certainly most reliable one consists in studying the specific features of the data, and then use any available tool to design a specialized method for the matching task. This approach will generally take advantage of language-specific (e.g. in (Freeman et al., 2006)) and domain-specific knowledge, of any external possible resources (e.g. names dictionaries, etc.), and of any information about the entities to process (especially their type: for example, there are differences between person names and organizations). In such an in-depth approach, human expertise is required in numerous ways.

The second approach is the *robust* one: we propose here to try to match any kind of NE, extracted from "real world" (potentially quite noisy) sources, without any kind of prior knowledge[1]. One looks for coreferent NE, whatever their type, source, language[2] or quality[3]. Such robust similarity methods may be useful for a lot of generic tasks, in which maximum accuracy is not the main criterion, or simply where the required resources are not available.

The orthographic similarity between strings is usually evaluated through some sort of string similarity measure. The literature on string comparison metrics is abundant, containing both general techniques and more linguistically motivated measures, see e.g. (Cohen et al., 2003) for a review. From a bird's eye view, these measures can be roughly sorted in two classes[4]:

- "Sequential character-based methods", which look for identical characters in similar positions. The most well known is certainly the Levenshtein edit distance, for which there exists a lot of variants/improvements and efficient algorithms (Navarro, 2001); the Jaro distance is also commonly used in record linkage problems (Winkler, 1999).
- "Bag-of-words methods", which are based on the number of common words between two strings, irrespective of their position. In this category fall very simple measures like the Jaccard similarity or overlap coefficient, or more elaborated ones like the Cosine similarity applied to TF-IDF weights. A related family of measures applies the same kinds of computation to "bag of (characters) n-grams" representation.

The application of these measures is relatively well documented in the database literature (see e.g. (Winkler, 1999)); however, when dealing with named entities found in text collections, it is less clear which measure(s) should be considered (see however (Freeman et al., 2006; Pouliquen et al., 2006)). Furthermore, most work on named entity matching has focused on morphological (formal) similarity. Yet, a major difference between the record linkage application and text applications is the availability of information regarding the context of occurrences of entities. We expect that this extra-information could help solve cases that are difficult for the morphological similarity measures; a similar idea has already been used for disambiguating

[1] In this kind of knowledge are included the need for hand-tuning parameters or defining language-specific heuristics.

[2] Actually we have only studied English and French (our approach is neither "multilingual", in the sense that it is not specific to multingual documents).

[3] In particular, this task clearly depends on the NE recognition step, which may introduce errors.

[4] We omit measures based on phonetic similarity such as Soundex, because they are language-specific and/or type-specific (person names), and do not fit for text collections.

homonyms (Pedersen et al., 2005; Pedersen and Kulkarni, 2007).

Our long-term goal is to build a system for automatically detecting coreferent entities using multiple string comparison measures, through machine learning techniques to select an optimal combinations of measures. This approach however presupposes the availability of hand-labeled data, stipulating which pairs of entities are positive (coreferent), and which are negative (non-coreferent). Such data is required (i) to provide an objective criterion for selecting the best combination, and (ii) to evaluate the performance of the whole system.

As a first step in that direction, we thus present and discuss in this paper a methodology for building, in a semi-automatic manner, such a hand-labeled data. This methodology assumes that the only source of information comes from the corpus: in particular, we will not use any gazetteer. We will also assume that the preliminary text processing tasks have been performed, including named entity recognition, providing us with the locations of these entities in the documents. Finally, we assume that computation time is not restricted, and that it is possible to compute all the possible pairwise comparisons. This assumption is clearly unrealistic for very large data collections and in that case, one should resort to the use of *blocking*[5] techniques. However, in the context of the small corpora we have considered, such computation is indeed feasible, and enables us to study matching results independent from the bias that this filtering step may introduce.

When building a gold standard for referent named entities, two simple minded ideas should be immediately disregarded: (i) labeling all the existing pairs is clearly beyond reach, for this would require to examine n^2 pairs of entities, where n typically ranges in the thousands; (ii) performing a random sampling in the set of pairs would also be of little help: a randomly chosen pair of entities is almost always negative. In order to recover as many positive pairs as possible, we adopted the following methodology: first, a battery of similarity measure was computed for all the pairs of entities; the top n matches for all measure were then examined and manually labeled. This allowed us to systematically compare the matches provided by each (type of) measure. This approach was successively applied on two different corpora: based on the outcome of our first experiment, we had to somewhat refine the labeling guidelines, and extend the automatic labeling tools.

This paper is organized as follows: in Section 2., we introduce the corpora, tools and guidelines that have been used to produce a golden set of matched entities. In Section 3., we provide and discuss the results of these experiments, before concluding in Section 4..

[5]In brief, blocking consists in clustering in a first step the whole set of entities, in such a way that potentially coreferent entities belong to the same cluster and that the number of entities in each cluster is minimal. This step is intended to avoid the global quadratic comparison over the whole set of pairs, needed otherwise. The question of blocking is itself very important in record matching problems (Bilenko et al., 2006).

2. Data, approach and experiments

2.1. Input data

The first corpus we used, called "Iran nuclear threat" (INT in short), is in English and was extracted from the NTI (*Nuclear Threat Initiative*) web site[6], which collects all public data related to nuclear threat. It mainly contains news, press articles and official reports obtained from various (international) sources. This corpus, limited to the 1991-2006 years, is 236,000 words long (1.6 Mio). It was chosen because

- it contains informations from various sources, a diversity that guarantees the existence of orthographic variations in named entities,
- it focuses on Iran and is thus bound to contain many transliterated names (from Persian or Arabic)

This data is slightly noisy, due to the variety of sources and/or extraction errors. We used GATE[7] as the named entities recognizer. Recognition errors are mainly truncated entities, over-tagged entities, and common nouns beginning with a capital letter. We restricted the set of entities only to those belonging to one of the three categories: locations, organizations and persons (as recognized by GATE). We obtained this way a set of 35,000 (occurrences of) entities. We finally decided to work only on the set of entities appearing at least twice, resulting in a set of 1,588 distinct entities accounting altogether for 33,147 occurrences.

Our second corpus, called "French speaking medias" (FSM in short), is a 856,000 words long corpus, extracted from a regular crawling of a set of French-speaking newspapers web sites during a short time-frame (in July 2007). The web sites were chosen based on the following criteria: geographic diversity, large volume of content, ease of access. Once again, we made sure to include a large number of web sites from North Africa, a potential source of transliterated Arabic names.

The extraction was performed by Pertimm[8]. The tagging of named entities in the corpus was then performed by Arisem[9], recognizing a total of 34,000 occurrences of entities recognized as locations, persons or organizations. Once again, the recognition step is noisy, but significantly less so than with the English corpus: less truncated or over-tagged entities, but slightly more false entities (mainly common nouns; the latter is easier to deal with than the former: for evaluation purposes, false entities have simply to be discarded). In the following, we will only work on the set of entities appearing at least twice, which yielded a unique set of 2,533 "real" entities, corresponding to 23,725 occurrences.

2.2. Methodology

Our string matching system is intended to test, evaluate, and compare as much as possible all available similarity measures. Overall, we experimented with 48 different measures, 20 of which where imported from existing

[6]http://www.nti.org
[7]http://gate.ac.uk
[8]http://www.pertimm.com
[9]http://www.arisem.com

open source packages: SimMetrics[10] by S. Chapman and SecondString[11] by W. Cohen, P. Ravikumar and S. Fienberg. Following (Christen, 2006), (Cohen et al., 2003), (Bilenko et al., 2003), we mainly considered the following measures[12]:

- *Sequential character-based:* Levenshtein, Jaro, Jaro-Winkler, Needleman-Wunch, Smith-Waterman and variants.
- *Bag of words:* Cosine, Jaccard, Overlap (simply using the number of common words between two strings), cosine with TF-IDF weighted vectors of words.
- *N-grams-characters based (for n=1,2,3):* Jaccard-type, cosine with TF-IDF weighted vectors of n-grams.
- *Combinations of measures:* Monge-Elkan, Soft-TFIDF (proposed in (Cohen et al., 2003)).
- *Context based:* this measure correspond to the Cosine of the TF-IDF vectors representing the context of two entities; context vectors contain all the occurrences of the words occurring within a fixed distance of each entity.

Given an annotated corpus, our system performs the following computations:

1. Read the NE data and the reference dataset (whenever available), select a subset of entities to process.
2. Compute the whole matrix of measures for all (selected) entities and measures[13]: each measure is applied to every pair of entities yielding $n \times (n-1)/2$ scores.
3. Manually tag top ranking pairs as positive or negative (optional).
4. For each measure, compute the k best pairs (for a predefined value of k). For several predefined values $m \leq k$, it is then possible to evaluate the individual performance of each similarity measure, using the traditional precision/recall/f-measure metrics. Additionally, it is possible to assess how each measure behaves with respect to parameters like length, number of words or frequence.
5. For every pair of measures, we finally compute the correlation coefficient and the number of common [positively labeled] pairs in the m best scores.

[10] `http://www.dcs.shef.ac.uk/~sam/simmetrics.html`

[11] `http://secondstring.sourceforge.net`

[12] A detailed description of these measures may be found on S. Champan's web page: `http://www.dcs.shef.ac.uk/~sam/simmetrics.html`.

[13] We do not distinguish entities by type (persons, locations, organizations), because type errors are rather frequent in both corpus, therefore comparing only entities having the same type would miss some positive pairs (for example, in our datasets different occurrences of the same NE are sometimes labeled with different types).

2.3. Semi-automatic labeling

As explained above, it would be very costly to manually label as match (positive) or non-match (negative) the whole set containing $n \times (n-1)/2$ pairs, for the observed values of n. A standard solution would be to label only a randomly chosen subset of pairs: in the special case of this task, this approach is ineffective, because of the disproportion between the number of positive and negative pairs. In fact our datasets only contain only respectively 0.06% (for INT) and 0.02% (for FSM) positive pairs. This is why we tried to find all the positive pairs, assuming that the remaining lot are negative. Practically, the labeling step was based only on the best pairs as identified by our set of measures. This is clearly a methodological bias (very roughly, measures are evaluated on the basis of their own predictions), but we hope to have kept the effects of this bias as low as possible. This is because the measures we used are quite diverse and do not assign good scores to the same pairs; therefore, for each measure, we expect that the potential misses (false negatives) will be matched by some other measure, thus allowing a fair evaluation of its performance. Basically this approach is close to the TREC pooling evaluation method (see e.g. (Voorhees and Harman, 1998)): the battery of measures acts as the different participating systems. Evaluation issues are further discussed in Section 3.2..

2.3.1. Labeling the INT

For the INT corpus, the labeling is based solely on the best pairs retrieved by the different measures. For each measure, our system provides the sorted set of the k best pairs, which were then proposed for human labeling in decreasing order. A minimal number of pairs is labeled for each measure (approximatly 1000), in order not to unbalance results between measures.

The guidelines we used for labeling this corpus are the following:

- *positive pairs:* two entities are considered matching if there is a "quite obvious" coreference link. Coreference is here interpreted in a rather loose sense:
  - if one of the entities is not correctly tagged (small truncation or containing too many words), they may be labeled positive provided they are clearly recognizable. Example: *"Bushehr Nuclear Plant"*, *"Completing Bushehr Nuclear Plant"*
  - in some slightly ambiguous cases, two entities are considered matching if the coreference link is highly probable. For example, *"US Senate foreign relation commission"*, *"Senate foreign relation commission"* is a positive pair because the corpus never talks about the *"Senate foreign relation commission"* of another country, even if such another commission may actually exist. Also, some cases of metonymy are considered positive, although this choice is certainly questionable: for instance, "Europe" and "Western Europe" are considered matching.
- *negative pairs:* two real (well formed) entities are labeled negative only if there is no doubt about their non-coreference.

- *"don't know"*[14] *pairs:* all other cases, including:
  - at least one entity is incomplete, not recognizable or ill-formed,
  - the coreference link is doubtful (potential homonymy, lack of knowledge/information from the corpus), semantic ambiguity (e.g. *"Foreign Ministry", "Russian Foreign Ministry"*).

The choice of a relatively loose definition for positive pairs was guided by the concern to label a maximum amount of positive data. The manual labeling eventually yielded 805 positive pairs, 1,877 negative pairs and 3,836 "don't know" pairs.

2.3.2. Labeling the FSM

For the French corpus, labeling was more elaborated: we used the n best pairs from each measure, but also added two new methods. The first one consists in trusting transitivity relationships: if entities A and B match and entities B and C match, then entities A and C match[15]. The second one, which is more time-consuming, is a new pass over the whole set of entities. For each entity e, the n closest entities e' according to m "good" measures were also proposed for a human annotator[16]. This provides a different (complementary) viewpoint than processing the global n best pairs: this way, some pairs that could not obtain a top ranking score (this is typically the case of short entities, which are systematically over-ranked by longest ones) have a chance to be matched. The guidelines used for labeling have also been improved, based on the experience gained on the first one:

- *positive pairs:* strict coreference, at least in the corpus. The main objective is to preserve transitivity, thus it is not possible to consider "approximative coreference matching".
- *negative pairs:* strict non-coreference.
- *uncertain pairs:* this class consists is all pairs that are rejected from the positive ones but nonetheless present an important link. Some examples are: *"ONU" (UN)* and *"Conseil de sécurité" (Security Council), "Russie" (Russia)* and *"Gouvernement russe" (Russian Government*).
- *eliminated entities:* all others, which consist mostly in ill-formed entitites, but also a few special ambiguous cases.

Compared with the first corpus, more time has been spent looking for possible matches in the set of entities. For example, a lot of acronyms were manually matched against their development[17] and several special cases like *"Quai d'Orsay" and "Ministère des affaires étrangères"*[18] were also addressed. Finally, the use of a supplementary processing pass allowed to label a handful of additional positive pairs (approximately a dozen among around 30,000). For all these reasons, we think that the probability for a positive pair not to be labeled is very low. We finally labeled 741 positive pairs, 32,348 negative pairs and 419 uncertain pairs. 745 entities were discarded as ill formed in the process.

3. Experiments and discussions

Performances are evaluated under the following hypotheses, in agreement with our manual labeling procedure (see above): any unlabeled pair is considered as a negative one, and any pair marked as "don't know" (or uncertain) is simply ignored.

3.1. Main observations

Overall, all measures proved to behave similarly on both corpora. Differences are nonetheless observed between the achieved performance, which are significantly worst in the case of French-speaking medias corpus. As explained above (see parts 2.3.), this is mainly due to the fact that our labeling guidelines were more strict with this second corpus.

Measures that seem to perform best are "bag of words" measures, which compute a score given the number of common (identical) words between the two strings. As expected, taking into account the IDF (Inverse Document Frequency) gives slightly better results, that is why Cosine computed over TF-IDF weighted vectors (of words) is globally the best mesure. This seems to indicate there is a pay-off in working directly with words (as opposed to characters, n-grams characters and/or positional parameters) when comparing named entities. It is indeed true that most named entities of interest, be they person or organization names, tend to correspond to morphologically complex units (title/function+first name+last name for persons, nominal groups for organizations). Yet, this result is not entirely expected, as the Cosine distance between entities is very sensitive to small orthographic differences.

In fact, it appears that in the subset of the more easily matched pairs (pairs that appear very often as one of the best scores with any measure), sequential character-based methods perform better. This subset mostly contains pairs of long strings that only differ by one or two characters. Therefore, these pairs will eventually be also matched by word-based methods, as they also contain more words than the average (they are long), and several of which are indeed identical. These pairs will thus be matched by any measure. The main problem with character-based methods is that they have a hard time sorting out the more difficult cases.

By contrast, characters n-grams measures, particularly for n=2,3, achieve an overall better level of performance. An examination the best ranking pairs for these measures reveals that they combine features from bag of words and

[14] This category is distinct from the (really) unlabeled pairs, because it does not contain any positive or negative pair.

[15] Similarly, if A and B match but A and C do not, then B and C do not match.

[16] In practice, we used $n = 3$ and $m = 4$.

[17] Although this kind of match is out of the scope of textual similarity measures, so we do not expect to catch them.

[18] *"Quai d'Orsay"* is the address where the French Ministry of Foreign Affairs is located, and is very often used as a metonym for the Ministry.

Table 1: Positive pairs by frequence

	INT	FSM
frequence ≥ 2	805	741
frequence ≥ 3	386	421
frequence ≥ 5	202	212
frequence ≥ 10	64	72

sequential character based methods: they catch minor differences more easily than bag of words measures, but have two drawbacks: firstly, as the other ones, they favour long strings (because probability to find common n-grams is higher). Secondly, they are sometimes "confused" by long strings containing similar n-grams in a different order, thus bringing a bit more false positive than bag of words measures.

Finally, the context-based measure is a very poor individual measure. As expected, good scores are obtained for entities which have an important semantic link. But this is not precise enough to match coreferent entities: typically, an organization may be matched with the person who is its main representative. A lot of other false positives are found, such as *"Israel"* and *"Palestine"*. However, the rare true positive found are interesting, because some of them could not be found by any textual measure (like acronyms and their development). This is why we plan to use the context measure in conjunction with other measures, hoping that in this case, it will prove more useful than used in isolation.

Overall, all the (good) measures tested tend to favour long strings: the average lengths in our corpora are respectively about 13 and 11 characters long (1.9 and 1.8 words long), whereas the average length among 500 best scores for all measures is respectively 15.4 and 13.1 characters long (2.1 words long for both). We also note that the average frequency of high ranking pairs is very high compared to the global average frequency. This may be due to the fact that very frequent entities are more likely to appear with variations (observing matched pairs corroborates this hypothesis).

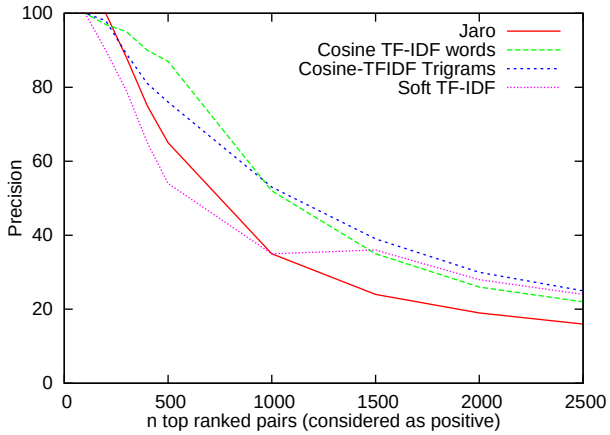
In our corpora, the most frequent sources of variation can be roughly classified as follows:

- Small typographical differences about spaces, diacritic signs, upper case letters. For example, in the FSM corpus *"Al Qaïda"* appears under 7 variations (with *i* or *ï*, with or without the hyphen, with or without uppercase *A*). These variations are easily captured by sequential character based or n-grams based methods.

- Omissions are very frequent in organization names, as in *"United States"* and *"United States of America"*, or in *"Conseil de Sécurité [de l'ONU / des Nations Unies]"* (*[UN / United Nations] Security Council*), where a PP modifier is omitted. Bag of words methods generally perform well on this kinds of pairs.

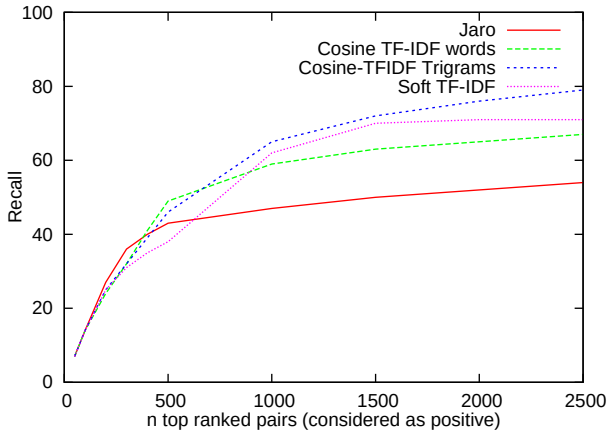
- Person names with or without the first name are also very frequent.

Figure 1: Precision (FSM)



Precision for four string comparison measures.

Figure 2: Recall (FSM)



Recall for four string comparison measures.

- Geographical orthographic variations may be more or less complex to identify, ranging from the simple pair *"Darfur"* and *"Darfour"* to the more challenging pair *"parc national d El Kala" / "parc naturel de la Calle"*.

Overall, all these variations are well taken care of, at least by one family of measures. More difficult cases occur when several sources of variations are combined, e.g. a change in a person name accompanied by the deletion of the first name as for the pair *"Lugovoï"* / *"Andreï Lougovoï"*.

Unsurprisingly, false positive pairs are entites that are orthographically similar but do not match, like *"ministère chinois des Affaires étrangères"* and *"Ministère russe des affaires étrangères"* (*Chinese/Russian Ministry of Foreign Affairs*) or *"South Africa"* and *"South America"*.

3.2. Discussion

The main pitfall in evaluating entities matching techniques in this framework is the disproportion between positive and negative data, together with the fact that it is (almost) im-

possible to label the whole data. As described in part 2.3., the method used to catch positive pairs depends on measures themselves. This means that there might remain some unlabeled positive pairs, which are wrongly counted as negative ones in the evaluation. This does not affect the computed precision, since enough pairs have been labeled among good scores for each measure. But recall should be interpreted with this potential bias in mind, since it depends on the number of false negative which may be underestimated.

We have tried to quantify this effect by manually searching the 2,533 unique entities in FSM for unlabeled positive pairs. As expected most of those found did not present textual similarity (otherwise they would eventually have been detected by similarity measures). Most of them were acronyms, but some other examples are also worth mentioning: *"M. Ban"* and *"Ban Ki Moon"*, *"aéroport Congonhas"* and *aéroport international de Sao Paolo"* (*Sao Paulo International Airport*), *"USA"* and *"États-Unis"* (*United States*). Under the hypothesis that we did not forget any pair, we can roughly express the probability that a positive pair remains undetected by our procedures is about 5%. A last note is in order: in all our experiments, we only considered those words that actually occurred at least twice: orthographic variations due to typos, which typically occur only once, are probably underestimated.

One of the questions we studied carefully concerns the length of entities. All (good) measures favour long strings, therefore it is possible that some pairs of short entities are missed. We have looked for best scores among short strings, in particular by filtering only entities containing only one or two words. We also studied how the distribution of the length of strings behaves with respect to the scores for several measures. Although this can not replace a systematic labeling, our observations suggest that there are simply less matching pairs with short entities, because possible textual variations are naturally proportional to the string length.

Finally, the case of uncertain pairs is also worth discussing. In our experiments, these were simply ignored; a fairer evaluation of name entity match should take them under consideration, using an intermediate status between positive and negative. For example, the pair *"ministère des Affaires étrangères"*, *"ministère français des Affaires étrangères"* (*"Ministry of foreign affairs"*, *"French Ministry of foreign affairs"*) is uncertain, although most occurrences of the general form concern the French Ministry. This question is related to another one: what is the limit for a pair to match ? Even if all occurrences of *"Ministry of foreign affairs"* in the corpus refer to the French one, should one consider this pair as a match or consider the question in a more general context: the latter viewpoint has the advantage to permit to accumulate knowledge (e.g. for large dynamic databases), contrary to the former.

4. Conclusion and future work

In this paper, we have proposed a methodology for semi-automatically labeling data in a NE matching problem, and studied the problems that arise from this methodology. We have shown that this task, which consists in finding coherent entities extracted from corpus, presents the following peculiarities:

- very small set of positive pairs compared to the whole set of possible pairs (0.02% and 0.06% in our corpora). This problem makes it hard to obtain a sufficient amount of labeled data, thus introducing potential evaluation issues.

- some string similarity measures perform well, but no unique (existing) measure seems able to capture the variety of observed phenomena. Taking only one individual measure to compare entities requires either to make a compromise between precision and recall performance or to rely to a post-processing human validation step (as used in a lot of real systems, such as (Pouliquen et al., 2006)).

As a side note, it is worth mentioning that most sources of variations are captured by at least one family of measures.

In the future, we therefore plan to investigate methods for combining several measures, in order to improve the overall matching performances. There are different ways to do so: the first one is to use supervised learning techniques, using the now available sets of labeled data. One may also try to build new measures that would be more suited to the NE matching problem, since most existing measures are simply *string* similarity measures. In particular, it seems especially relevant to investigate unsupervised learning, or at leat semi-supervised learning techniques (for example, asking user to label only a limited number of chosen pairs).

Acknowledgements

This work has been funded by the National Project Cap Digital - Infom@gic. We thank Loïs Rigouste (Pertimm) and Nicolas Dessaigne and Aurélie Migeotte (Arisem) for providing us with the annotated French corpus.

5. References

Mikhail Bilenko, Raymond J. Mooney, William W. Cohen, Pradeep Ravikumar, and Stephen E. Fienberg. 2003. Adaptive name matching in information integration. *IEEE Intelligent Systems*, 18(5):16–23.

Mikhail Bilenko, Beena Kamath, and Raymond J. Mooney. 2006. Adaptive blocking: Learning to scale up record linkage. In *ICDM*, pages 87–96. IEEE Computer Society.

Peter Christen. 2006. A comparison of personal name matching: Techniques and practical issues. Technical Report TR-CS-06-02, Department of Computer Science, The Australian National University, Canberra 0200 ACT, Australia, September.

William W. Cohen, Pradeep Ravikumar, and Stephen E. Fienberg. 2003. A comparison of string distance metrics for name-matching tasks. In Subbarao Kambhampati and Craig A. Knoblock, editors, *Proceedings of IJCAI-03 Workshop on Information Integration on the Web (IIWeb-03), August 9-10, 2003, Acapulco, Mexico*, pages 73–78.

Andrew Freeman, Sherri L. Condon, and Christopher Ackerman. 2006. Cross linguistic name matching in english and arabic. In Robert C. Moore, Jeff A. Bilmes, Jennifer Chu-Carroll, and Mark Sanderson, editors, *HLT-NAACL*. The Association for Computational Linguistics.

Gonzalo Navarro. 2001. A guided tour to approximate string matching. *ACM Comput. Surv.*, 33(1):31–88.

Ted Pedersen and Anagha Kulkarni. 2007. Unsupervised discrimination of person names in web contexts. In *Proceedings of the Eighth International Conference on Intelligent Text Processing and Computational Linguistics*, Mexico City, february.

Ted Pedersen, Amruta Purandare, and Anagha Kulkarni. 2005. Name discrimination by clustering similar contexts. In Alexander F. Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing, 6th International Conference, CICLing 2005, Mexico City, Mexico, February 13-19, 2005, Proceedings*, volume 3406 of *LNCS*, pages 226–237. Springer.

Bruno Pouliquen, Ralf Steinberger, Camelia Ignat, Irina Temnikova, Anna Widiger, Wajdi Zaghouani, and Jan Zizka. 2006. Multilingual person name recognition and transliteration. *CORELA - Cognition, Representation, Langage.*, September 11.

Ralf Steinberger, Bruno Pouliquen, and Camelia Ignat. 2006. Navigating multilingual news collections using automatically extracted information.

Ellen M. Voorhees and Donna Harman. 1998. The text retrieval conferences (trecs). In *Proceedings of a workshop on held at Baltimore, Maryland*, pages 241–273, Morristown, NJ, USA. Association for Computational Linguistics.

W. E. Winkler. 1999. The state of record linkage and current research problems. Technical Report RR99/04, US Bureau of the Census.