Is this NE tagger getting old?

Cristina Mota, Ralph Grishman

L2F (INESC-ID) & IST & NYU, New York University Computer Science Department
Rua Alves Redol 9 1000-029 Lisboa Portugal, New York NY 10003 USA
cmota@ist.utl.pt, grishman@cs.nyu.edu

Abstract

This paper focuses on the influence of changing the text time frame on the performance of a named entity tagger. We followed a twofold approach to investigate this subject: on the one hand, we analyzed a corpus that spans 8 years, and, on the other hand, we assessed the performance of a name tagger trained and tested on that corpus. We created 8 samples from the corpus, each drawn from the articles for a particular year. In terms of corpus analysis, we calculated the corpus similarity and names shared between samples. To see the effect on tagger performance, we implemented a semi-supervised name tagger based on co-training; then, we trained and tested our tagger on those samples. We observed that corpus similarity, names shared between samples, and tagger performance all decay as the time gap between the samples increases. Furthermore, we observed that the corpus similarity and names shared correlate with the tagger F-measure. These results show that named entity recognition systems may become obsolete in a short period of time.

1. Introduction

A recent survey by (Nadeau et al., 2007), covering the last 15 years of research in named entity recognition (NER), shows that the field has been growing significantly, in terms of the number of languages processed, textual genres and domains covered. As the authors describe, the availability of large amounts of data motivated researchers to gradually abandon the early hand-crafted rule systems, and adopt (supervised and semi-supervised) machine learning techniques.

The performance of these systems is usually measured by testing the system on unseen texts, which will be compared to a gold standard or key, to assess precision, recall and F-measure in the identification and classification of the named entities (NE). In general, the training and test conditions are similar, i.e., the system is trained with texts that are comparable in terms of genre, topic and language to the test texts; otherwise the performance will most certainly decrease. What if we keep the same genre, topic and language, but change the time frame of test texts? Do texts vary over time in a way that will affect the performance of a name tagger? Such issues are important in selecting the training data for a tagger.

In this paper we show that as the time gap between training and test texts increases (i) the similarity between the two texts decreases, (ii) the name lists of those texts overlap less; and (iii) the performance of an NE tagger also decays. Furthermore, we will show that the results of the corpus analysis and the system performance correlate over time.

The paper is structured as follows: we begin by a brief discussion on how time has been dealt with in natural language processing; then we characterize the task we are assessing and the corpus used in our experiments; in sections 4. and 5., we present, respectively, the methodology to analyze the corpus, and the name tagger architecture; finally, we show the results of our experiments over time and conclude, pointing to future directions.

2. The influence of time in NLP

Time in natural language processing can be viewed through two different perspectives: as a relevant object that can be processed or as a system variable that conditions its performance. The former has been by far more explored, and consists in recognizing and relating temporal information in texts (e.g., identification of temporal expressions, identification of when an event took place, ordering events chronologically and novelty detection).

We are more interested in the second perspective that sees time as a system variable, i.e., there are changes in texts over time (either due to language changes or to topic and sub-topic shifts) that affect the performance of natural language systems and, hence, systems should be conceived time-aware.

The majority of works we found address this question in an indirect way by recognizing that linguistic resources are never complete. For instance, (Stevenson and Gaizauskas, 2000) explore NE annotated corpora to automatically update gazetteers and (Fairon and Courtois, 2000) use online newspapers to enlarge dictionaries of common words. This option follows the suggestion *"more data is better data"* (Church and Mercer, 1993), focusing on the idea of collecting more data, which indeed enlarge the previous resources, but without a concern for selecting relevant data that would increase even more the resources or would produce tailored resources for analysing a specific corpus. Some authors, however, show that simply adding more data in the training stage is not enough: (Ji and Grishman, 2006) show that enlarging the size of training data doesn't always yield the best results, and they stress the need of carefully selecting appropriate data to bootstrap a name tagger; (Atterer and Schutze, 2006), in a study about disambiguating prepositional phrase and relative clause attachment, also argue that the size of the unannotated corpus had little effect on the performance not only due to the noisiness of the statistics extracted from the unannotated corpus, but also because the corpora were from distinct sources and time periods.

The idea of having similar temporal data for training and testing NLP systems is also shared by other authors. (Gale

and Church, 1994) when comparing different probability estimators for English bigrams, split a one year corpus into training and test by assigning each bigram starting at an even-numbered word to the former, and those starting at odd-numbered words to the latter. In this way, they were aiming at obtaining closer training and test samples, avoiding temporal bias (they had observed measurable differences over the period of one month). (Church, 1995) studied the correlation between word variants (singular/plural, adjective/adverb, lower case/upper case) by comparing the correlation estimates obtained in one year of the Associated Press corpus with the estimates obtained on different years of the same news agency. He observed that the estimates were highly reliable over a period of five years (from 1988 to 1992), but since the estimates degraded over time, he suggested that in order to predict the correlation between two word forms in one year one should use the estimates of the previous year instead of the estimates obtained ten years before.

As in this latter work, we trained our name tagger in one time frame and then tested the tagger with texts within other time frames. Moreover, we also compared the texts over time in different ways, relating the results of the corpus analysis with the results of the tagger performance.

3. Experimental Conditions

In this section we begin by characterizing the task our name tagger was developed for, and the data used in our experiments.

3.1. Task definition

The first evaluation for named entity recognition in Portuguese, HAREM, took place in 2004 (Cardoso and Santos, 2007). We adopted the HAREM NE annotation scheme, but simplified the classification task to approximate the MUC named entity task (Grishman and Sundheim, 1995): (i) we are only interested in proper names of people, organizations and locations; (ii) assignment of type and morphological attributes is not considered; (iii) the classification gives preference to the classification of the base form rather than to the name function, as in the "form over function" approach (Johannessen et al., 2004). This last decision means that, for instance, Portugal is classified as location in both of these contexts:

Portugal is located in Europe

Portugal voted against the proposal

In HAREM, these references should be classified as location and organization, respectively. For a thorough comparison between MUC and HAREM tasks cf. (Seco, 2007).

3.2. Data description

We conducted the experiments on samples of the Politics articles of the Portuguese corpus CETEMPúblico (Rocha and Santos, 2000). The corpus has 180 million words and spans 8 years, from 1991 to 1998, divided into semesters. The original newspaper articles were fragmented by the corpus builders into extracts (typically 2 paragraphs each, which in Politics resulted in about 5 sentences per extract), and then randomly shuffled to comprise the final corpus. This means that within a semester, two adjacent extracts most likely don't belong to the same article and that they are not sorted temporally.

The Politics section of the corpus corresponds to about one quarter of the corpus. From this section, we created a golden collection (key) of 16 texts by manually annotating with NE the first 400 extracts in each semester; from those we used the first 192 extracts to collect the tagger seeds and the remaining 208 as test sets; the following 7856 extracts of each semester are also used by the NE tagger as unlabeled data. We grouped these subsets by year.

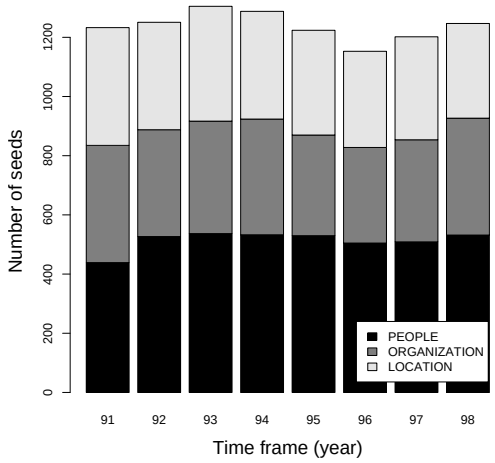


Figure 1: Seeds distribution by topic and category

The tagger seeds correspond to the set of different names obtained after collecting the first 2500 name instances of each year in the golden collection. Figure 1 shows the distribution of seeds by category in each time frame. Even though, there are more seeds classified as people (about 40% of the different seeds are people names, and the remaining names are equally distributed by organizations and locations), people names are less frequent: in average, each name occurs about 1.6 times, whereas organization and location names occur about 2.29 and 2.34, respectively. Hence, in terms of name instances the category distribution is 32.2%, 34.81% and 32.99% for people, organization and location, respectively.

From the 7856 unannotated extracts of each semester, the tagger uses the first 82456 unlabeled names and corresponding surrounding context as pairs of unlabeled examples to bootstrap. Those names occur in one of the contexts mentioned in section 5.

4. Corpus comparison

The corpus was compared over time through two different perspectives. On the one hand, we compared the raw corpus, i.e., the corpus without any annotations besides the original corpus structural tags, simply by comparing word frequency lists. On the other hand, we compared the name lists of the golden collection that we manually annotated and the name lists extracted by the name tagger. We begin

by describing the corpus similarity approach and then the name list overlap metrics.

4.1. Corpus homogeneity/similarity

In order to measure corpus similarity over time, we adopted the method proposed by (Kilgarriff, 2001) to compare language varieties. Given two corpora A and B with the same number of words, the author measures the similarity between A and B, $similarity(A, B)$, by applying the following algorithm:

- Split corpus A and B into k slices each
- Repeat m times:
  - Randomly allocate $\frac{k}{2}$ slices from A to A_i and $\frac{k}{2}$ slices from B to B_i
  - Construct word frequency lists for A_i and B_i
  - Compute distance between word frequency lists of A_i and B_i for the n most frequent words of the joint corpus (A_i+B_i)
- Output mean and standard deviation of the distances obtained in all experiments $i = 1 \ldots m$

After measuring $similarity(A, B)$, one must also compute $similarity(A, A)$ and $similarity(B, B)$, which give the within-corpus similarity, i.e. corpus homogeneity, of corpora A and B, respectively.

The author argues that the similarity between two corpora can only be properly interpreted if one compares it to the corpus homogeneity of each corpus: for instance, if the similarity between the two corpora is small but their homogeneity is high, then one may conclude the two corpora belong to different language varieties. (Baayen, 2001) also advocates a similar position stating that to understand the importance of an intertextual difference one should account for the intratextual variability of the text characteristic being analyzed.

In his study, Kilgarriff compared different distance metrics by applying the algorithm to Known-Similarity Corpora (KSC). As χ^2 by degrees of freedom (CBDF) performed the best on the KSC, we used this distance in our experiments. The remaining parameters of the algorithm, n and k, were established empirically and equal 2000 and 10, respectively.

Instead of randomly allocating the slices from both A and B, we iterate over all combinations of 5 slices drawn from a partition of 10, select those slices of A for A_i, and select the complementary-numbered slices of B for B_i. Furthermore, if a set of 5 indices for A and 5 for B has already been used to compute a distance, we exclude the case where these two sets are interchanged. This corresponds to running $\frac{C_5^{10}}{2}$ experiments, i.e., 126 experiments. In this way, we guarantee that in the homogeneity experiments: (i) we don't compare the same halves more than once; (ii) we don't compare overlapped halves, i.e., we don't assign a slice with the same index to both halves. The consequence of doing this is that $similarity(A, B) \neq similarity(B, A)$. Since the values are always very close, we didn't average the two to obtain a single value for the similarity between A and B. This decision is reasonable because, as we will see, all corpora are highly homogeneous, and also because, in any case, we made all possible comparisons within the time interval.

4.2. Name list overlaps

Even though the previous approach can be extended to compare name lists, we wanted to have an idea of how much overlap existed between name lists over time, i.e., how many names were shared between texts within different time frames.

As a measure of overlap between names occurring in one reference text and each of the remaining texts, we used Jacquard's coefficient. This metric calculates the number of name types that occur in both texts relative to the total number of different names in the joint texts. As an alternative, which gives more weight to frequent names, we also measured the name token overlap. Given texts A and B, and the frequencies of name i in set A, $f_A(i)$, and in set B, $f_B(i)$, the name type overlap and the name token overlap correspond to equations (1) and (2), respectively.

$$type\_overlap = \frac{|T_A \cap T_B|}{|T_A| + |T_B| - |T_A \cap T_B|} \quad (1)$$

$$token\_overlap = \frac{\sum_{i=1}^{N} min(f_A(i), f_b(i))}{\sum_{i=1}^{N} max(f_A(i), f_B(i))} \quad (2)$$

These metrics were used to compare the names occurring in the golden collection and also the names extracted by the name tagger.

5. NE tagger

In this section we describe the system we implemented and its performance on each time frame.

5.1. Tagger description

We adopted and modified the semi-supervised name tagger based on DL-cotraining proposed by (Collins and Singer, 1999) for two main reasons:

1. It is a simple semi-supervised method, requiring just a few labeled seeds;
2. It performs well when compared to supervised methods.

Given that we have limited manually annotated data, it offered a good tradeoff between available training data and quality of the result.

This method separates the identification and classification stages. In terms of named entity recognition, only the classification involves learning. In the identification stage, the text is parsed and the NEs occurring in particular syntactic contexts are collected along with their surrounding context, constituting pairs of name and context. In the classifier training stage, these unlabeled pairs will be classified alternately using spelling rules (based on the names) and contextual rules, and from the examples that were labeled, contextual or spelling rules will be inferred, respectively. The algorithm starts by using the spelling seeds. At each

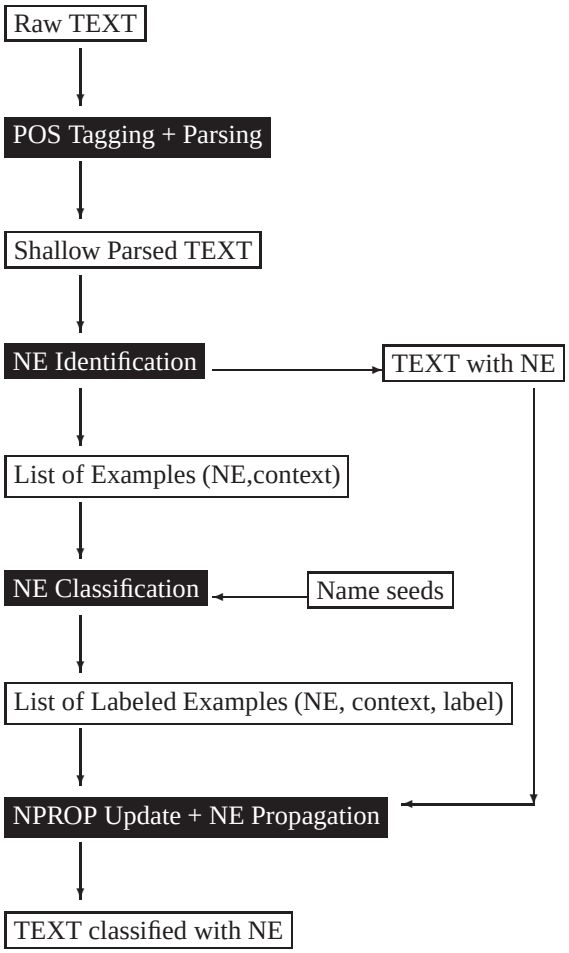


Figure 2: NE tagger architecture

step the most frequent features which co-occur with labeled examples of each NE type are examined; if the strength, $h(x, y)$ (an estimate of the conditional probability $P(y|x)$ of seeing label y when x is in the example feature set) exceeds a threshold (0.95), this feature/label pair is added to the new set of rules, forming a decision list. In the last step all possible rules (both spelling and contextual) are added to the final decision list. This final decision list is used to label a test set obtained by applying the same sequence of operations as for the training set: POS tagging, parsing, NE identification, and NE feature extraction.

In our implementation, we modified the identification stage. We used NooJ (Silberztein, 2004) to do local parsing and extract the NEs occurring in the following contexts: head of noun phrase, complement of noun phrase, left or right context of a verb, coordination, and age context; in the classification stage, we used five of the features proposed by the authors (namely, full_string, contains, allcaps1, context_type and context) and also used the length of the NE. Moreover, we integrated an additional propagation stage to boost recall, i.e., the classified names will be used to recognize in the text other instances of the same names occurring in contexts that were not identified by the first stage.

The overall architecture of our name tagger is depicted in Figure 2.

5.2. Tagger evaluation

Before analyzing the performance over time, by training and testing the system within different time frames, we evaluated our system in order to understand how well it performed within each time frame.

(Collins and Singer, 1999) evaluated their tagger by measuring accuracy and clean accuracy[1] on a sub-set of 1000 examples they manually labeled, and randomly selected from the 90000 examples they had identified, obtaining 83.3% and 91.3%, respectively. We evaluated our tagger by measuring the precision, recall and F-measure regarding each text in the golden collection.[2]

We compared the performance of our tagger to the performance of a baseline tagger. Since the classification algorithm uses seeds to bootstrap, our baseline consisted in simply looking up the seeds and assigning to the name in the test text the most frequent label of that name in the seed set. The baseline was also applied to each text in the golden collection.

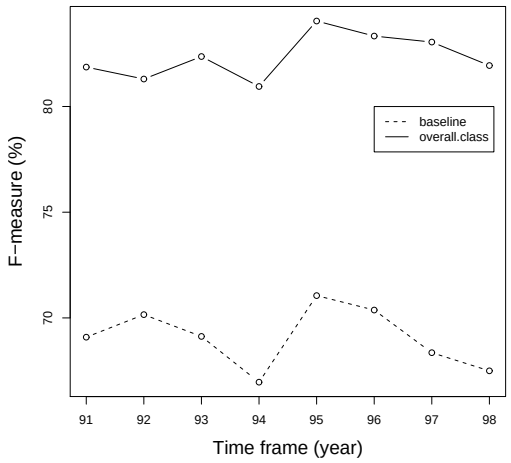


Figure 3: Classification F-measure for co-training and baseline taggers

As can be seen in Figure 3, the baseline classification F-measure is lower than 71.06% for all texts within the golden collection, averaging 69.08%, whereas our tagger classification is above 80.95% for all texts, achieving in average 82.36%.

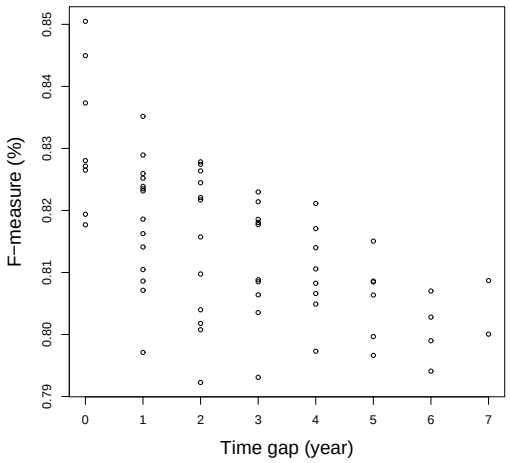
6. Results

In the following experiments, we denote by S_i, U_i and T_i, respectively, the seed, unlabeled and test texts, by $NE(t)$ the names extracted by the tagger from text t, and by ($S_{i=91...98}$, $U_{i=91...98}$, $T_{j=91...98}$) a training-test configuration.

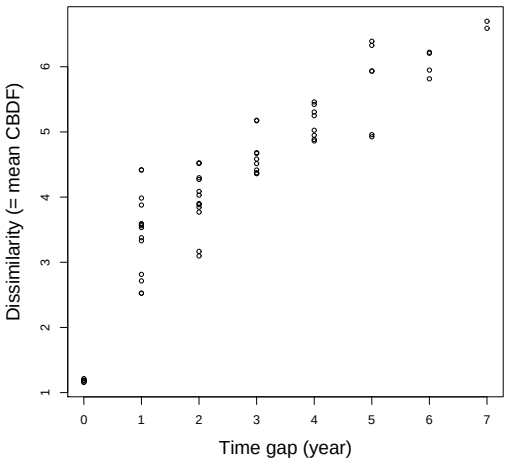
For each text (indexed by k) in the time interval from 91 to 98, we ran the NE tagger with the configuration (S_k, U_k,

[1]Contrarily to accuracy, clean accuracy doesn't take into account examples that were incorrectly identified, i.e., which do not belong to one of the categories people, organization or location

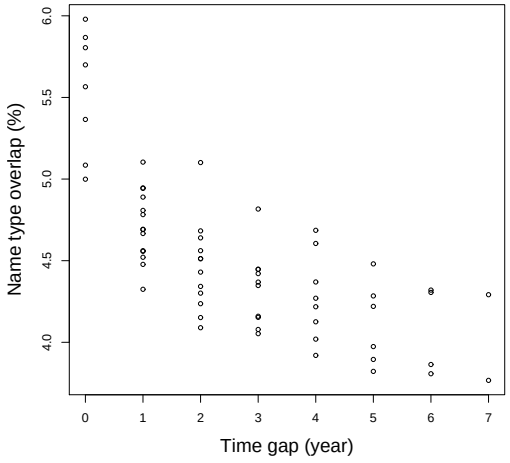
[2]We used the scoring programs of the HAREM evaluation (Santos et al., 2007).



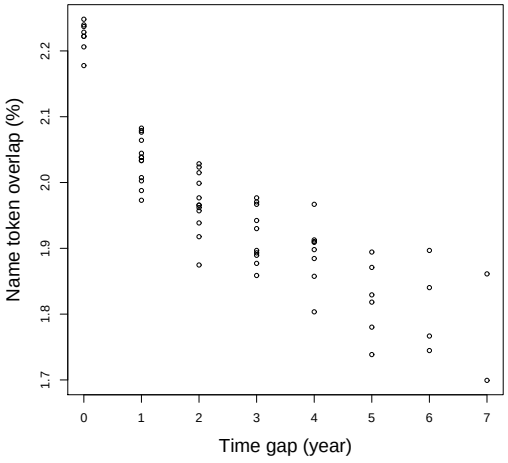
(a) F-measure for ($S_{k=91...98}$, $U_{k=91...98}$, $T_{j=91...98}$)



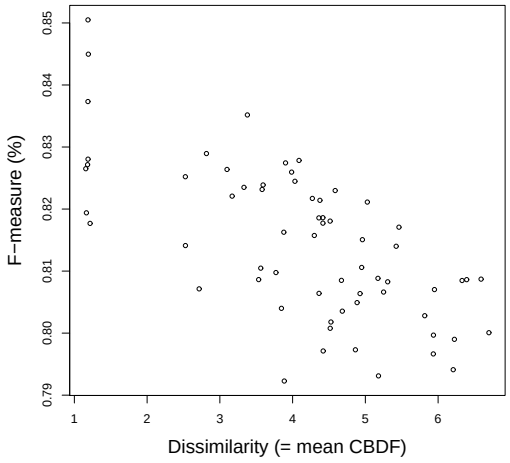
(b) Dissimilarity for ($U_{k=91...98}$, $U_{j=91...98}$)



(c) Name type overlap for ($NE(U_{k=91...98})$, $NE(T_{j=91...98})$)



(d) Name token overlap for ($NE(U_{k=91...98})$, $NE(T_{j=91...98})$)



(e) F-measure compared to dissimilarity

Figure 4: Corpus and system analysis over time from 1991 to 1998

$T_{j=91...98}$), i.e., we trained the system with data from one year and then tested it with data within each year in the time interval. This procedure corresponds to 64 different experiments[3], we calculated:

- the F-measure and the classification accuracy for each $T_{j=91...98}$;
- the corpus similarity between U_k and $U_{j=91...98}$;
- the type and token name overlap between $NE(U_k)$ and $NE(T_{j=91...98})$.

Figure 4 illustrates the most significant results.
As can be seen in Figure 4(a), as the time gap between (S_k, U_k) and T_j increases, the tagger F-measure shows a tendency to decay. Notice, for instance, that when the texts are from the same year (time gap = 0), the F-measure ranges approximately from 82% to 85%, but when the texts are 5 years apart the F-measure ranges from about 79% to 82%.
For the corpus similarity between U_k and U_j, represented in Figure 4(b) in terms of dissimilarity (i.e., mean CBDF distance), the decreasing tendency is more evident: the homogeneity for all the texts (i.e., within-corpus distance, corresponding to a time gap of 0) is very close to 1, but just by increasing the time gap to one year, the dissimilarity ranges from 2.5 to 4.5, and at a distance of five years it ranges from 4.7 to almost 6.5.
Figure 4(c) shows that the name type overlap also decreases as the time gap between U_k and T_j increases: for instance, when the texts are within the same time frame, the overlap varies between 5% and 6%, whereas at a distance of 5 years it varies between 3.5% and 4.5%. The name token overlap (Figure 4(d)) shows a similar decay: within the same year, the overlap varied between 4.2% and 4.4%, and at distance of 5 years between 3.2% and 3.7%. The overlap values are so low because $NE(U_k)$ is much larger than $NE(T_{j=91...98})$.
It is also worth mentioning that the decreasing tendency of these properties as the time gap increases doesn't seem to flatten. This suggests that if the time interval was larger, the decreasing tendency would continue.
Finally, figure 4(e) shows that there is an inverse association between dissimilarity and F-measure: for higher values of dissimilarity we obtain lower performance values.

7. Concluding remarks and future directions

We showed that within a period of 8 years, the corpus similarity and name overlaps tend to decrease as the two corpora become more temporally distant, and that the performance of a co-training based NE tagger trained and tested on those texts reflects the dissimilarity between them by showing a decay in performance as we increase the time gap between the training data and the test data. Furthermore, we showed that there is an association between the results of the corpus analysis and the tagger performance.

[3]We notice that for a time gap of 0 there are 8 experiments, and for each time gap t there are $2 \times (Y - t)$ experiments, where: $0 < t < 8$ and Y is the number of years in the interval.

Given these preliminary results, one future direction we intend to pursue is to analyze the NE surrounding contexts to verify if they also tend to overlap less over time. The other goal will be to investigate how we can avoid the performance decay. The immediate answer would be adding more data, and not just any data, but temporally relevant data.

8. Acknowledgments

The first author's research work was funded by Fundação para a Ciência e a Tecnologia through a doctoral scholarship (ref.: SFRH/BD/3237/2000). We are also grateful to Adam Kilgarriff for his prompt support when we first attempt to implement his method.

9. References

Michaela Atterer and Hinrich Schutze. 2006. The effect of corpus size in combining supervised and unsupervised training for disambiguation. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 25–32, Sydney, Australia, July. Association for Computational Linguistics.

R. Harald Baayen. 2001. *Word Frequency Distributions*, volume 18 of *Text, Speech and Language Technology*. Springer.

Nuno Cardoso and Diana Santos. 2007. Directivas para a identificação e classificação semântica na colecção dourada do harem. In Diana Santos and Nuno Cardoso, editors, *HAREM, a primeira avaliação conjunta de sistemas de reconhecimento de entidades mencionadas para português: Documentação e actas do encontro*. Linguateca.

Kenneth W Church and Robert L Mercer. 1993. Introduction to the special issue on computational linguistics using large corpora. *Computational Linguistics*, 19(1):1–24.

Kenneth Ward Church. 1995. One term or two? In *SIGIR '95: Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 310–318, New York, NY, USA. ACM Press.

Michael Collins and Yoram Singer. 1999. Unsupervised models for named entity classification. In *Proceedings of the Joint SIGDAT Conference on EMNLP*.

Cédrick Fairon and Blandine Courtois. 2000. Les corpus dynamiques et glossanet. extension de la couverture lexicale des dictionnaires Électroniques anglais. In *Actes des 5es Journées Internationales d'Analyse Statistique des Données Textuelles (JADT 2000)*, Lausanne.

William A. Gale and Kenneth W. Church. 1994. What's wrong with adding one? In N. Oostdijk and P. de Haan, editors, *Corpus-Based Research into Languge: In honour of Jan Aarts*, pages 189–200. Rodolpi, Amsterdam.

Ralph Grishman and Beth Sundheim. 1995. Design of the MUC-6 Evaluation. In *Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference held in Columbia, Maryland, November 6-8, 1995*, Los Altos, Ca. Morgan Kaufmann.

Heng Ji and Ralph Grishman. 2006. Data selection in semi-supervised learning for name tagging. In *Proceedings of the Workshop on Information Extraction Beyond The Document*, pages 48–55, Sydney, Australia, July. Association for Computational Linguistics.

Janne Bondi Johannessen, Kristin Hagen, Asne Haaland, Andra Bjork Jansdottir, Anders Naklestad, Dimitris Kokkinakis, Paul Meurer, Eckhard Bick, and Dorte Haltrup. 2004. Named entity recognition for the mainland scandinavian languages. *Literary and Linguistic Computing*, 20(1):91–102.

Adam Kilgarriff. 2001. Comparing corpora. *International Journal of Corpus Linguistics*, 1(6):1–37.

Nadeau, David, Sekine, and Satoshi. 2007. A survey of named entity recognition and classification. *Linguisticae Investigationes*, 30(1):3–26, January.

Paulo Rocha and Diana Santos. 2000. Cetempúblico: Um corpus de grandes dimensões de linguagem jornalística portuguesa. In Maria das Graças Volpe Nunes, editor, *Actas do V Encontro para o processamento computacional da língua portuguesa escrita e falada PROPOR 2000*, pages 131–140, Atibaia, São Paulo, Brasil.

Diana Santos, Nuno Cardoso, and Nuno Seco. 2007. Avaliação no harem: Métodos e medidas. In Diana Santos and Nuno Cardoso, editors, *HAREM, a primeira avaliação conjunta de sistemas de reconhecimento de entidades mencionadas para português: Documentação e actas do encontro*. Linguateca.

Nuno Seco. 2007. MUC vs HAREM: a contrastive perspective. In Diana Santos and Nuno Cardoso, editors, *HAREM, a primeira avaliação conjunta de sistemas de reconhecimento de entidades mencionadas para português: Documentação e actas do encontro*. Linguateca.

Max Silberztein. 2004. Nooj: A cooperative, object-oriented architecture for nlp. In *INTEX pour la Linguistique et le traitement automatique des langues*, Cahiers de la MSH Ledoux. Presses Universitaires de Franche-Comté.

Mark Stevenson and Robert Gaizauskas. 2000. Using corpus-derived name lists for named entity recognition. In *Proceedings of 6th Applied Natural Language Processing Conference and 1st Meeting of the North American Chapter of the Association for Computational Lingusitics*, pages 290–295, Seattle.