Knowledge discovery with CRF-based clustering of named entities without a priori classes

Vincent Claveau, Abir Ncibi

Knowledge discovery with CRF-based clustering of named entities without a priori classes

Vincent Claveau and Abir Ncibi

IRISA-CNRS and INRIA-IRISA
Campus de Beaulieu, 35042 Rennes, France
vincent.claveau@irisa.fr, abir.ncibi@inria.fr

Abstract. Knowledge discovery aims at bringing out coherent groups of entities. It is usually based on clustering which necessitates defining a notion of similarity between the relevant entities. In this paper, we propose to divert a supervised machine learning technique (namely Conditional Random Fields, widely used for supervised labeling tasks) in order to calculate, indirectly and without supervision, similarities among text sequences. Our approach consists in generating artificial labeling problems on the data to reveal regularities between entities through their labeling. We describe how this framework can be implemented and experiment it on two information extraction/discovery tasks. The results demonstrate the usefulness of this unsupervised approach, and open many avenues for defining similarities for complex representations of textual data.

1 Introduction

Labeling sequences are tasks of particular interest for NLP (part-of-speech tagging, semantic annotation, information extraction, etc.). Many tools have been proposed, but in recent years, the Conditional Random Fields (CRF [1]) have emerged as the most effective for many applications. These models are supervised machine learning: examples of sequences with their labels are required.

The work presented in this paper is placed in a different context in which the goal is to bring out information from these sequences. So, we fit in a task of knowledge discovery in which supervision is not applicable: the aim is to discover how the data can be grouped into categories that make sense rather than providing these categories from expert knowledge. Therefore, these discovery tasks are based most often on clustering [2,3,4]; the crucial question is how to calculate the similarity between two interesting entities. In this paper, we propose to divert CRF by producing fake labeling problems in order to make appear entities that are regularly labeled the same way. Of these regularities is then built a notion of similarity between entities, which is thus defined by extension and not by intention.

On the application point of view, in addition to the use for knowledge discovery, the similarities obtained by our approach or the clusters produced can be used upstream of supervised tasks:

- it can be used to reduce the cost of data annotation. It is indeed easier to label a cluster than annotate a text instance by instance.
- it can help to identify classes difficult to discriminate, or on the contrary exhibit classes whose instances are very diverse. It then makes it possible to adapt the supervised classification task by changing the set of labels.

In the remainder of this article, we position our work in the state-of-the-art and briefly present CRF by introducing some useful concepts for the rest of the article. We then describe in Section 3 the principle of our discovery approach using supervised ML technique in an unsupervised mode for discovery tasks. Two experiments of this approach are then proposed in Sections 4 and 5, before presenting conclusions and future work in the last section.

2 Related Work

Many NLP tasks are nowadays considered as supervised ML problems: they suppose the existence of a set of pre-defined classes, and, of course, examples belonging to these classes. Yet, several studies have proposed moving to a non-supervised framework. Some of these studies are not, strictly speaking, about non-supervision but rather about semi-supervision since their goal is to limit the number of sequences to be annotated. It is particularly the case for the recognition of named entities; indeed, many studies rely on external knowledge bases (e.g. Wikipedia) or extraction rules given by an expert as a bootstrap [5,6,7]. Let us also mention the work on Part-of-Speech tagging without annotated data by HMM [8], Bayesian training [9], integer programming [10] or other approaches [11,12]. Similar work has been proposed, along with other formal frameworks for named entities [13,14]. More recently, entity linking tasks have been explored [15], their goal is to link a string mention in a document to an existing entry/category in a database. In all cases, the perspective of these studies is different than ours as they do not adopt a knowledge discovery setting: they are all based on a *tagset* already established.

The framework that we adopt in this paper is different: we aim at making categories emerge from unannotated data. Unlike previously cited work, we do not make any *a priori* on the possible label set. The task is

therefore a clustering one, in which similar elements from the sequences should be grouped, as it was done for example by [4] for some named entities. Clustering words is not a new task in itself, but it relies on the definition of a representation for words (typically a context vector) and a measure of distance (or similarity, typically a cosine). Our approach aims to use the discriminative power of ML tools to provide a more effective measure of similarity. The goal is therefore to turn these techniques from supervised into unsupervised for determining the similarity between any two elements of sequences. In this paper, we report experiments using CRF, which has proved its efficiency for numerous supervised tasks [16,17,18,19, inter alia], but it is worth noting that the whole approach can be applied with other ML methods.

This way of diverting supervised machine learning techniques to bring out similarities in complex unlabeled data has been used for data mining. It was demonstrated as very useful for propositional data (i.e. described by feature-value pairs) for which defining a similarity was difficult (non numeric attributes, bias of a definition *ex nihilo*). Different ML methods have been used in this framework, including Decision Trees and Random Forests [20,21,22]. The approach consists in generating a large number of artificial learning problems, with generated synthetic data that are mixed with the real data, and then in stating what data are classified together regularly. Our approach fits into this framework, but exploits the peculiarities of CRF in order to take into account the sequential nature of textual data.

CRFs [1] are undirected graphical models that represent the probability distribution of annotations (or labels, or tags) y conditionally on observations x. More precisely, in the case of sequences like sentences, the conditional probability $P(y|x)$ is defined through the weighted sum of feature functions f and g. They are usually binary functions satisfying a certain combination of labels and attributes describing the observations and applied at each sequence position: f functions characterize the local relations between the current label in position i and observations; functions g characterize the transitions between the nodes of the graph, that is, between each pair of labels at position i and $i-1$, and the sequence of observations. These functions are defined by the user according to his knowledge of the application; their weights reflect their importance to determine the class. Learning CRF consists in estimating these weights (the vector of weights is noted θ hereafter) from training data. Indeed, from N labeled sequences, the vector θ is searched as the one that maximizes the log-likelihood of the model on these labeled sequences. In practice, this

optimization problem is solved by using quasi-Newton algorithms, like L-BFGS [23]. After the learning phase, the application of CRF to new data consists in finding the most probable sequence of labels y^* given a (previously unseen) sequence of observations x. As for other stochastic methods, y^* is generally obtained with a Viterbi algorithm.

3 Principles of the unsupervised model

This section describes the principle of our approach. An overview is first given through an algorithm depicting the whole process. We then detail some crucial points, as well as more insights about the practical use of this method.

3.1 General principle

As we explained above, the main idea of this approach is to derive a distance (or similarity) from repeated classifications of two objects for random learning tasks. The more often objects are detected as belonging to a same class, the closer they are supposed to be. The approach chiefly relies on the fact that CRF will make it possible to exhibit similarity between words by assigning them repeatedly identical labels in varied learning conditions. As for bagging [24], the learning process is repeated several times with different settings in order to change the learning bias. For this, several random choices are being implemented at each iteration; they concern:

- the sequences used for learning;
- the labels (number and distribution);
- the feature functions describing words.

These supervised learning tasks on artificial problems should confer, with their variety, important properties of the similarity obtained. It naturally handles complex descriptions (for instance the various attributes of the current word, the word neighbours); it operates a selection of variables by construction, and thus takes into account descriptor redundancies or ignores those of poor quality, and is robust to outliers.

Algorithm 1 gives an overview of the process. The sequential classification with CRF is repeated many times with varying data labels (the ω_i are fake classes) and learning parameters (feature functions, training set $\mathcal{E}_{\text{train}}$). The model is then applied to the data not used as training set, called 'out-of-bag' ($\mathcal{E}_{\text{OoB}}$). Pair of words (x_i, x_j) receiving same labels are memorized, and these *co-labelings*, kept in the matrix $\mathcal{M}_{\text{co-label}}$.

Algorithm 1 Clustering by CRF

input: $\mathcal{E}_{\text{total}}$: non labeled sequences
for great number of iterations **do**
  $\mathcal{E}_{\text{train}}$, $\mathcal{E}_{\text{OoB}} \leftarrow$ Divide($\mathcal{E}_{\text{total}}$)
  Randomly choose labels y_i among $\omega_1 ... \omega_L$ for sequences in $\mathcal{E}_{\text{train}}$
  Randomly generate a set of feature functions f and g
  Infer: $\theta \leftarrow$ L-BFGS($\mathcal{E}_{\text{train}}$,$y$,$f$,$g$)
  Apply: $y^* = \arg\max_y p_{(\theta,f,g)}(y|x)$ for each $x \in \mathcal{E}_{\text{OoB}}$
  for all classe ω_l among $\omega_1 ... \omega_L$ **do**
    for all pair x_i, x_j of $\mathcal{E}_{\text{OoB}}$ such that $y_i^* = y_j^* = \omega_l$ **do**
      $\mathcal{M}_{\text{co-label}}(x_i, x_j) += \text{weight}(x_i, x_j, \omega_l)$
$\mathcal{M}_{sim} \leftarrow$ Transform($\mathcal{M}_{\text{co-label}}$)
return Clustering($\mathcal{M}_{sim}$)

From this matrix, similarities between each pairs can be derived (possibly with a simple normalization of the co-label counts) and then used by the clustering algorithm.

3.2 Random learning

Of course, as we have already pointed out, the important role of randomness does not prevent the user to control the task through different bias. This is reflected for example by the provision of rich descriptions of words: part-of-speech tags of the sequences, semantic information of the words... This is also reflected in the definition of the set of feature functions from which the algorithm can draw the functions f and g at each iteration. In the experiments reported below, these functions are those usually used for Named Entity recognition : word-form, part-of-speech and upper or lower case status from the current word, the 3 preceding and the 3 following, bigrams built from these features... Concerning the sets $\mathcal{E}_{\text{train}}$ and $\mathcal{E}_{\text{OoB}}$, at each iteration, 5% sentences are randomly chosen as the training set and the remaining serves as application set.

3.3 About random labels

In many applications, the task of clustering is only useful for a subset of the words/phrases in the texts. In this context, it is very common to use BIO type labels that can model multi-word entities (B indicates an entity beginning, I the continuity, O is for words outside entities). Table 1 presents an example of artificial sequences derived from the data used in Section 5:

x	l’	audience	entre	nicolas	sarkozy	et	maître	wade	...
	DET	NC	PREP	NP	NP	COO	NC	NP	...
y	O	O	O	B-fake140	I-fake140	O	B-fake25	B-fake3	...

Table 1. Example of sequence with observations (words, parts-of-speech) and fake labels

This external knowledge is part of the essential biases needed to control the process of unsupervised learning and to ensure that it applies to the specific needs of the user. But it is important to note that this knowledge about which entities have to be considered is not the same than the one that we aim to discover via clustering. In the first case, it consists in spotting the entities while in the second case, it consists in making emerge classes of entities without a priori.
It is possible in this latter case to assume that we know how to delimit the interesting entities in the sequences; this is the assumption made in several studies on the classification of named entities [13,14,4]. It is also, of course, possible to consider this problem as a learning problem itself for which the user must provide some examples. In both cases, this requires expertise, provided either by intention (objective criteria to define entities) or extension (cf. Subsection 5.2). Each of the experiments reported below adopts one of these two cases.

The choice of the number of the fake labels, as well as their distribution, is important (yet, it has to be underline that the number of labels does not directly impact on the number of clusters that will be eventually generated). A very high number of fake labels may produce a model difficult to infer (CRF complexity is very dependent on the number of labels), and may also result, when applied to $\mathcal{E}_{\text{OoB}}$, in data with few entities sharing the same labels. On the opposite side, if too few random labels are used for the inference step, the model obtained may be not discriminative enough and thus may produce fortuitous co-labeling in $\mathcal{E}_{\text{OoB}}$. Of course, all of this is also dependent on the many other parameters of the inference. For instance, the feature functions may allow or not over-fitting, and thus possibly prevent or favour co-labeling of entities. The size of $\mathcal{E}_{\text{train}}$, and more specifically the number of entities that may receive the same fake label is also important : if, on a systematic basis, many training entities from probably different classes share the same fake label, the model will tend to be not discriminating enough.

In order to correctly take into account these phenomena, it would be necessary to characterize, before the labeling step and ideally before the inference step, the tendency of the model to discriminate entities enough

or not. Unfortunately, such an a priori criterion is difficult to formalize. Instead, we use a simple a posteriori regularization: the co-labeling of two entities is considered as more informative if few entities have received this label. This is implemented as a weight function used when updating the $\mathcal{M}_{\text{co-label}}$ matrix. In practice, in the experiments reported below, this weighting function is defined as: $\text{weight}(x_i, x_j, \omega_l) = \frac{1}{|\{x_k | y_k = \omega_l\}|}$ and the number of labels is randomly chosen between 10 and 50 at each iteration.

For some discovery tasks, according to the particular knowledge available for them, it is also possible to bias the distribution of the random labels in the training set. For instance, if one knows that every occurrence of an entity necessarily belongs to the same class, it is important to implement this constraint in the training step. The experiment detailed in Section 4 falls within this framework.

3.4 Clustering

The final step of clustering can be implemented in various ways using different techniques and tools. The famous k-means algorithm requires centroid calculations during the process; this is of course not suitable for our non-metric space, but its variant k-medoids, which uses an object as a representative of a cluster, does not require other similarity/distance measures than those provided by $\mathcal{M}_{sim}$.

Let us underline here that in discovery tasks, the number of clusters to be produced is of course unknown. For our part, in the experiments presented in Sections 4 and 5, another clustering technique is considered, namely Markov Clustering (MCL). This technique was originally developed for partitioning large graphs [25]. Its advantage over k-medoids is that it does not require to set a priori the number of clusters expected, and it also avoids the problem of initialization of these clusters. We therefore consider only our objects (words or other entities of the sequences) as vertices of a graph whose edges are valued based on the similarity contained in $\mathcal{M}_{sim}$.

3.5 Operational aspects

The iterative process proposed in this paper is obviously expensive but easily parallelizable. In the experiments reported below, the number of iterations was set to 1,000; with such a high number of iterations, the results obtained are stable and can be reproduced despite the several random steps of the algorithm. The main sources of cost in terms of

Time	Report
80mn	Zigic donne quelques frayeurs à Gallas et consorts en contrôlant un ballon chaud à gauche des 16 mètres au devant du Gunner. Le Valencian se trompe dans son contrôle et la France peut souffler.
82mn	Changement opéré par Raymond Domenech avec l'entrée d'Alou Diarra à la place de Sidney Govou, pour les dernières minutes. Une manière de colmater les brêches actuelles ?

Table 2. Excerpt of a minute-by-minute football report

computation time are learning CRF models and the application of these models to label data. The complexity of these steps depends on many parameters, including the size of the training sample, the variety of observations (x), the number of random classes (ω), the attributes considered (feature functions f and g)... To minimize the impact of this cost, we use an implementation of CRF WAPITI that optimizes standard inference algorithms [26].

4 Experimental validation: classification of proper names

For this first experiment, we consider the problem and data of [4]. The goal is to bring out the various classes of proper names in football (soccer) summaries. More specifically, in their experiments, the authors have attempted to classify names at the corpus level: all occurrences of a same proper name are considered to belong to a unique entity and thus to a unique class. Therefore, in this dataset, entities are not considered as polysemous; even if that point is debatable, we adopt here the same framework than [4].

4.1 Task and data

The corpus is composed of minute-by-minute match reports in French, taken from various websites. Important events of almost every minute of a match are described (see Table 2): player replacements, fouls, goals...

These data have been manually annotated by experts according to classes defined to meet specific application requirements [27]. These annotations constitute a ground truth: it defines what could be interesting classes, and it associates each proper name to a class (see Figure 1). Note that the classes are very unbalanced with a large *player* class.

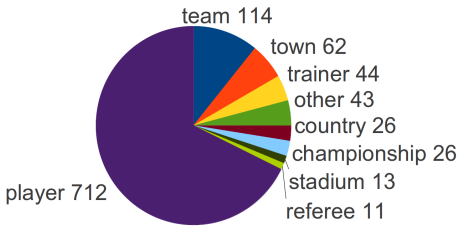

Fig. 1. Distribution of proper names in the football ground truth (number of unique names)

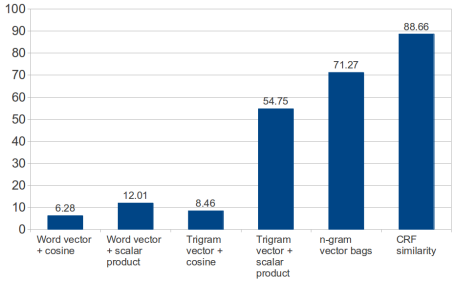

Fig. 2. Clustering evaluation (ARI %); football dataset

4.2 Performance measures

Since our task of discovery relies on a final stage of clustering, it is evaluated as such. Evaluation of clustering tasks is always difficult: evaluation through external criteria requires to have a reference clustering (ground truth) whose relevance can always be discussed, but the internal criteria (e.g., a measure of cohesion of clusters) are known not to be reliable [28]. We therefore prefer the external evaluation: the clustering obtained by our process is compared with the ground truth produced by experts.

To do this, various evaluation metrics have been proposed, such as purity or *Rand Index* [29]. These measures, however, have a low discriminating power and tend to be overly optimistic when the ground truth contains classes of very different sizes [30]. We therefore prefer the *Adjusted Rand Index* (ARI), a version of the Rand index taking into account the agreement by chance, which has become a standard measure for clustering evaluation and is known to be robust. As secondary measures, we also indicate, when possible, V-measure, normalized mutual information and adjusted purity. Their study and definitions can be found in [31].

4.3 Implementation and results

To test our clustering method by CRF, the corpus was part-of-speech tagged; we use BIO annotation scheme to generate the fake labels. In this particular application, we take the assumption of [13,14]: entities to categorize are supposed to be known and defined. In practice, it means that the random labels are only generated for these entities; the other words in the corpus receive the label 'O'. Functions f and g are those conventionally used in information extraction: functions f bind the current label y_i

to observations (word-form and part-of-speech of the current word in x_i, word-form and part-of-speech in x_{i-1}, x_{i-2} x_{i+1}, x_{i+2}, or combinations of these features), the functions g bind two successive labels (y_{i-1}, y_i). Since the task here is to classify proper nouns at the corpus level and not at the occurrence level, we force two occurrences of a same name to have the same label when generating random labels (step 4 of the algorithm). However, since the CRF models annotate at the occurrence level, the matrix $\mathcal{M}_{\text{co-label}}$ keeps track of the occurrence classifications. The transformation step (step 11) transforms this matrix into a similarity matrix $\mathcal{M}_{sim}$ of proper names in the corpus by summing the rows and columns concerning the different instances of the same names.

The results of our approach are given in Fig. 2 in terms of ARI (percentage, 0 is a random clustering or a all-in-one clustering, and 100 is for a clustering identical to the ground truth). For comparison purposes, we report the results of [4]; these were obtained using vector description of the contexts of each entity, either as a single vector, or as bags of vectors, with suited similarity functions. The clustering step is performed with the same algorithm MCL than for our system. MCL has a parameter called inflation rate that affects indirectly the number of clusters produced. For a fair comparison, the results reported for each method are those for which this setting is optimal for the evaluation measure ARI. In these experiments, it corresponds to 12 clusters with the CRF-based similarity, and 11 for the n-gram bag-of-vectors.

These results emphasize the interest of our approach compared to more standard representations and similarities. The few differences between the clusters formed by our approach and the ground truth classes focus on the class *other*. This class contains the names of individuals appearing in various contexts (personality giving the kickoff, appearing in the audience...), with too few regularity to allow CRF, no more than other methods, to succeed at bringing out a similarity. It is worth noting that the density of entity in this corpus, and the fact that any entity is very likely to appear often in various contexts makes the corpus and the discovery task particular. It may explain why some errors reported by [4] as recurrent are not made by the clustering by CRF. For example, the vector methods tend to confuse the names of cities and names of players, as they often appear close to each other and therefore share the same contexts. These mistakes are not made by the CRF approach, for which the built-in consideration of the sequentiality in the labeling process (word order and label order are taken into account) help to distinguish between these two classes.

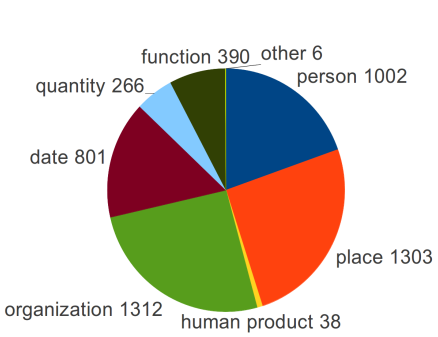

Fig. 3. Distribution of data in ESTER2 ground truth (number of occurrences)

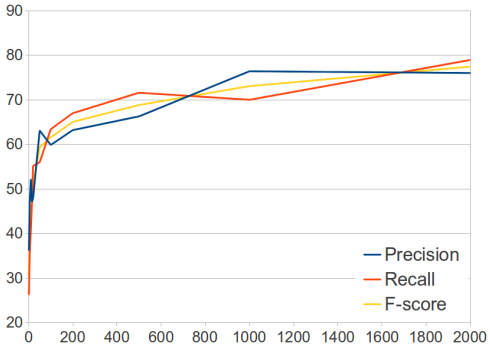

Fig. 4. Performance of entity detection according to the number of annotated training sequences

5 Experimental validation for information discovery

In order to assess the validity on another type of entity discovery task, this section presents further experiments on a news corpus, with a different definition of what are the interesting entities.

5.1 Task and data

The data are from the ESTER2 evaluation campaign [32]. They consist of 150 hours of radio recorded between 1999 and 2003 from various sources (France Inter, Radio Classique, Africa 1...). These broadcasts have been transcribed and then annotated for named entities according to 8 categories: people, functions, places, organizations, times, human products, quantity, and a *other* category.

Unlike the previous dataset, entities are annotated at the occurrence level; so, the entity `Paris` can be annotated as a place or organization depending on the context. In the experiments reported below, only the *dev* part of the ESTER2 dataset; it was transcribed manually, but respects the particularities speech recognition systems: the text has no punctuation or capitalization, which makes the named entity recognition task more challenging than for well-formed written texts. Here again, the manual annotation will serve as a ground truth for our discovery task; its characteristics are given in Figure 3.

5.2 Entity identification

Although it is possible to adopt the same framework as above and assume that the entities to classify are known and defined, here we use a more realistic framework: a small portion of the data is annotated by an expert who defines the entities of interest (but without assigning any class). These data will serve as a first step to learn how to retrieve entities before trying to cluster them. It is therefore a supervised task with two classes (interesting entity or not), for which we use CRF in its traditional way.

Figure 4 shows the detection results, depending on the number of sequences (phrases) used for learning. The performance is evaluated in terms of precision, recall and F-score. It appears that it is possible to retrieve the named entities from relatively few training sentences with good results (compared with the published results on close tasks on this dataset [32]).

5.3 Evaluation of clustering

The experimental framework is the same as in Section 4.3, except that the classification is done here at the occurrence level. The transformation of $\mathcal{M}_{\text{co-label}}$ in $\mathcal{M}_{\text{sim}}$ is therefore just a normalization. Entities considered are those identified by the previous step with 2,000 annotated sequences. The results, measured in terms of ARI, normalized mutual information, V-measure and adjusted purity are shown in Figure 5. As previously, we present the results of clustering techniques on the same data using more conventional similarities comparing the context and the entities (with the exception of the bags-of-bags approach that cannot apply to classification at the occurrence level). In order to assess the interest of our approach to handle complex representation of the data, we also add the results obtained by our CRF approach taking only into account the word-forms (no other features such as PoS).

On this task, the advantage of our approach is clear. Taking into account the sequentiality appears as very important: results with n-grams are indeed better than single words for the contextual vector representations, and our approach based on CRF, which take into account more naturally this sequential aspect, are even better. It is also worth noting that using the word-forms only yields slightly better results than ngrams; it underlines the interest of our approach compared to standard ones, even when using the same set of features, but it also emphasizes the benefit of

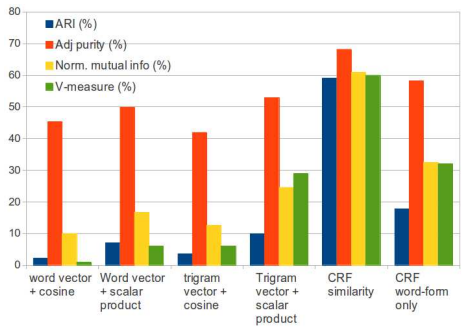

Fig. 5. Clustering evaluation; ESTER2 dataset

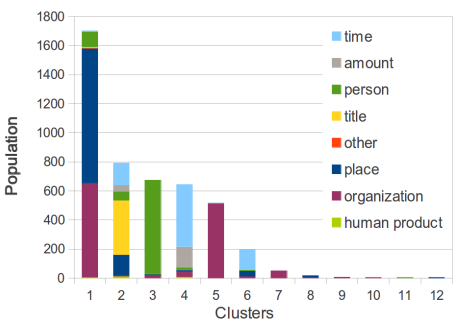

Fig. 6. Confusion between clusters and ground-truth classes

using complex representation (including PoS for instance), that are easily handled by our approach.

Clusters obtained by our approach, however, are not identical to those of the ground truth as one can see in Fig 6. Indeed, some clusters bring down the results by grouping entities belonging to distinct classes of ground truth. This is the case for 'organization' and 'place', which is a common mistake caused by polysemous names of country or town. This is also the case for 'time' and 'amount' which are difficult to distinguish without additional knowledge. Indeed, in the absence of other information than the form of words and parts-of-speech, it seems impossible to distinguish entities such as '`last four days`' (time) vs. '`on the last fifteen kilometres`' (amount).

6 Conclusion and perspectives

Solving fake learning problems with a ML technique helps to bring out similarities between textual elements. This similarity is making the most of the richness of description that the ML method allows (typically parts-of-speech, sequential information...). This defines a similarity in a non-metric space that is expected to be robust due to repeated random choices in the inference process. The use of CRF, successfully used for many (supervised) tasks, appears as an obvious choice, but of course, the same principles may be applied with other ML methods (HMM, MaxEnt, random forests, SVM)...

Evaluations conducted on two information extraction tasks[1] highlight the interest of the approach; although we are well aware of the limits of the evaluation of a discovery task which requires the establishment of a ground truth, which is what we want to avoid by using discovery techniques. It should also be noted that there is no machine learning without bias, even more when dealing with unsupervised learning [33]. These biases represent the knowledge of the user and help define the problem. The provision of information on entities to consider, the description of sequences and the definition of feature functions are pieces of information allowing the user to control the discovery task on its object of study.

Several improvements and perspectives are possible as a result of this work. From a technical point of view, the step transforming co-occurrences into similarities, which is a simple normalization in our experiences, could be deepened. Using other functions (such as those used to identify complex terms: mutual information, Jaccard, log-likelihood, χ^2...) to obtain more reliable similarities is foreseen. It may help to overcome the weakness of our clustering algorithm that can merge two clusters only because a few entities are strongly connected with many other nodes. Several variations concerning this clustering step may also be considered. It is for example possible to use hierarchical clustering algorithms. It is also possible to directly use the similarities between the words for other tasks, such as information retrieval, or for smoothing language models... From a practical point of view, it would be interesting to obtain an explicit definition of the similarity by recovering λ_i and μ_i along with their corresponding functions f and g. This would makes it possible to apply the similarity function to new texts without repeating the costly stages of learning.

References

1. Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: International Conference on Machine Learning (ICML). (2001)
2. Wang, W., Besançon, R., Ferret, O., Grau, B.: Filtering and clustering relations for unsupervised information extraction in open domain. In: Proceedings of the 20th ACM international Conference on Information and Knowledge Management (CIKM), Glasgow, Scotland, UK (2011) 1405–1414
3. Wang, W., Romaric Besan c., Ferret, O., Grau, B.: Evaluation of unsupervised information extraction. In: Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12), Istanbul, Turquie (2012)

[1] For replicability concerns, the football dataset is available from [27], ESTER2 from ELRA, WAPITI is available on `wapiti.limsi.fr`

4. Ebadat, A.R., Claveau, V., Sébillot, P.: Semantic clustering using bag-of-bag-of-features. In: Actes de le 9e conférence en recherche d'information et applications, CORIA 2012, Bordeaux, France (2012)
5. Kozareva, Z.: Bootstrapping named entity recognition with automatically generated gazetteer lists. In: Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop, Trento, Italy (April 2006) 15–21
6. Kazama, J., Torisawa, K.: Exploiting wikipedia as external knowledge for named entity recognition. In: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Prague, Association for Computational Linguistics (June 2007) 698–707
7. Wenhui Liao, S.V.: A simple semi-supervised algorithm for named entity recognition. In: Proceedings of the NAACL HLT Workshop on Semi-supervised Learning for Natural Language Processing, Boulder, Colorad, Association for Computational Linguistics (2009) 58–65
8. Merialdo, B.: Tagging english text with a probabilistic model. Computational Linguistics **20** (1994) 155–171
9. Richard, D., Benoit, F.: Semi-supervised part-of-speech tagging in speech applications. In: Interspeech 2010, Makuhari (Japan) (26-30 september 2010)
10. Ravi, S., Knight, K.: Minimized models for unsupervised part-of-speech tagging. In: Proceedings of ACL-IJCNLP 2009. (2009) 504–512
11. Smith, N., Eisner, J.: Contrastive estimation: Training log-linear models on unlabeled data. In: Proceedings of ACL. (2005)
12. Goldwater, S., Griffiths, T.L.: A fully bayesian approach to unsupervised part-of-speech tagging. In: Proceedings of the ACL. (2007)
13. Collins, M., Singer, Y.: Unsupervised models for named entity classification. In: Proceedings of Empirical Methods for Natural Language Processing (EMNLP) conference. (1999)
14. Elsner, M., Charniak, E., Johnson, M.: Structured generative models for unsupervised named-entity clustering. In: Proceedings of the Conference on Human Language Technology and North American chapter of the Association for Computational Linguistics (HLT-NAACL 2009), Boulder, Colorado (2009)
15. Ji, H., Grishman, R.: Knowledge base population: Successful approaches and challenges. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1. HLT '11, Stroudsburg, PA, USA, Association for Computational Linguistics (2011) 1148–1158
16. Wang, T., Li, J., Diao, Q., Wei Hu, Y.Z., Dulong, C.: Semantic event detection using conditional random fields. In: IEEE Conference on Computer Vision and Pattern Recognition Workshop (CVPRW '06). (2006)
17. Pranjal, A., Delip, R., Balaraman, R.: Part of speech tagging and chunking with hmm and crf. In: Proceedings of NLP Association of India (NLPAI) Machine Learning Contest. (2006)
18. Constant, M., Tellier, I., Duchier, D., Dupont, Y., Sigogne, A., Billot, S.: Intégrer des connaissances liguistiques dans un CRF : Application à l'apprentissage d'un segmenteur-étiqueteur du français. In: Traitement Automatique du Langage Naturel (TALN'11), Montpellier, France (2011)
19. Raymond, C., Fayolle, J.: Reconnaissance robuste d'entités nommées sur de la parole transcrite automatiquement. In: Actes de la conférence Traitement Automatique des Langues Naturelles, Montréal, Canada (2010)

20. Liu, B., Xia, Y., , Yu, P.: Cltree-clustering through decision tree construction. Technical report, IBM Research (2000)
21. Hastie, T., Tibshirani, R., Friedman, J.H.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction. New York: Springer (2001)
22. Shi, T., Horvath, S.: Unsupervised learning with random forest predictors. Journal of Computational and Graphical Statistics **15**(1) (2005) 118–138
23. Schraudolph, N.N., Yu, J., Günter, S.: A stochastic quasi-Newton method for online convex optimization. In: Proceedings of 11th International Conference on Artificial Intelligence and Statistics. Volume 2 of Workshop and Conference Proceedings., San Juan, Puerto Rico (2007) 436–443
24. Breiman, L.: Bagging predictors. Machine Learning **24**(2) (1996) 123–140
25. van Dongen, S.: Graph Clustering by Flow Simulation. Thèse de doctorat, Université d'Utrecht (2000)
26. Lavergne, T., Cappé, O., Yvon, F.: Practical very large scale CRFs. In: Proceedings the 48th Annual Meeting of the Association for Computational Linguistics (ACL), Association for Computational Linguistics (July 2010) 504–513
27. Fort, K., Claveau, V.: Annotating football matches: influence of the source medium on manual annotation. In: Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12), Istanbul, Turquie (2012)
28. Manning, C., Raghavan, P., Schütze, H.: Introduction to information retrieval. Cambridge University Press (2008)
29. Rand, W.M.: Objective criteria for the evaluation of clustering methods. Journal of the American Statistical Association **66**(336) (1971) pp. 846–850
30. Nguyen Xuan Vinh, Julien Epps, J.B.: Information theoretic measures for clusterings comparison. Journal of Machine Learning Research (2010)
31. Hubert, L., Arabie, P.: Comparing partitions. Journal of Classification **2**(1) (1985) 193–218
32. Gravier, G., Bonastre, J.F., Geoffrois, E., Galliano, S., Tait, K.M., Choukri, K.: ESTER, une campagne d'évaluation des systèmes d'indexation automatique. In: Actes des Journées d'Étude sur la Parole, JEP, Atelier ESTER2. (2005)
33. Mitchell, T.M.: The need for biases in learning generalizations. Rutgers Computer Science Department Technical Report CBM-TR-117, May, 1980. Reprinted in Readings in Machine Learning (1990)