Expanding a Gazetteer-Based Approach for Geo-Parsing Disease Alerts

Mikaela Keller MIKAELA.KELLER@CHILDRENS.HARVARD.EDU
John S. Brownstein JOHN.BROWNSTEIN@CHILDRENS.HARVARD.EDU
Clark C. Freifeld CLARK.FREIFELD@CHILDRENS.HARVARD.EDU
Children's Hospital Informatics Program at the Harvard-MIT Division of Health Sciences and Technology, 300 Longwood Ave., Boston, MA 02115 USA

Abstract

Discovering in a text the geographic references it may contain, is a task that human readers perform using both their lexical and contextual knowledge. Using a gazetteer to label such targeted references in a dataset, this paper proposes an approach to learning the context in which they appear and by this means extending the prior knowledge encoded in the gazetteer. The present work was carried in the particular framework of a system for disease outbreak alerts detection and geo-indexing.

1. Introduction

When presented in a text, with a phrase that is out of his vocabulary, a human reader would most likely be able to guess whether this phrase refers to a geographic location or not. This reader would infer the semantic role of the phrase with a certain accuracy, because he has a prior knowledge on the syntactic context on which geographic references appear, maybe also on their particular character distribution or on the fact that they generally begin with a capital letter, etc. There have been a number of approaches, exploiting this kind of prior knowledge to named entity recognition and more generally to information extraction problems (see *eg* (Tjong Kim Sang & De Meulder, 2003; Carreras & Màrquez, 2005)). They often rely on complex feature sets to represent the words and on heavily annotated datasets to account for the human experience.

HealthMap (Brownstein & Freifeld, 2007; Freifeld et al., 2008) is a system that automatically monitors disease outbreak alerts in news media from all around the world. It queries news aggregators such as Google News, but also news sources curated by experts, for relevant reports. It filters the retrieved documents into several taxonomies and provides on its website, `www.HealthMap.org`, a geographic and by-disease display of the ongoing alerts. Its operating mechanism can be naturally decomposed in a number of easily identifiable Information Retrieval and Natural Language Processing tasks, such as document retrieval, document categorization, information extraction, etc. In the present work we are interested in a sub-task of the last phase of the information processing scheme: the geographic parsing ("geo-parsing") (Woodruff & Plaunt, 1994) of a disease outbreak alert or the extraction from one such textual document of the related geographic information needed for the precise mapping into a world map.

So far, HealthMap assigns incoming alerts to a low resolution geographic description such as its country, and in some cases its immediately lower geographic designation (for the USA and Canada, it would provide for example the state). The system uses a rule-based approach relying on a purposely crafted gazetteer, which was built incrementally by adding relevant geographic phrases extracted from the specific kind of news report intended for mapping. The approach consists roughly in a look-up tree algorithm which tries to find a match between the sequences of words in the alert and the sequences of words in the entries of the gazetteer. It also implements a set of rules which use the position of the phrase in the alert to decide whether or not the phrase is related to the reported disease.

The gazetteer contains around 4000 key phrases, some of which refers to geographic locations with several resolution levels (from hospitals' to countries'), some are negation phrases ($\approx$ 500 phrases, *eg Brazil nut* or *turkey flock* are not considered location references) as well as phrases that are specific to the kind of data processed (*Center for Disease Control*, *Swedish health officials*, etc.).

HealthMap is interested in developing a higher resolution in the geographic assignments outside of those contained in the gazetteer. The question we would like to answer is whether we can use the prior knowledge encoded in the gazetteer to expand the system capabilities in the geographic parsing of the alerts.

Appearing in *Proceedings of the 25^{th} International Conference on Machine Learning*, Helsinki, Finland, 2008.

2. Our Approach

The basic idea behind our approach is to have a dataset of alerts tagged with the gazetteer-based algorithm as well as with more general linguistic knowledge (*eg* part-of-speech tags, etc.), and then to use this dataset with tags partially hidden to learn a generalization of the parsing. In the toy example of Fig. 1, a sentence is enhanced with its corresponding part-of-speech tags and gazetteer-based geo-parsing tags (the blue rectangle). In order to learn a generalization of the geo-parsing, the same sentence would be used in our training dataset with the specific identity of the word *New Caledonia*, hidden, but its part-of-speech, preserved.

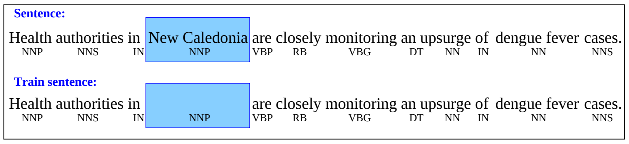
Sentence:
Health authorities in New Caledonia are closely monitoring an upsurge of dengue fever cases.
NNP NNS IN NNP VBP RB VBG DT NN IN NN NNS
Train sentence:
Health authorities in are closely monitoring an upsurge of dengue fever cases.
NNP NNS IN NNP VBP RB VBG DT NN IN NN NNS

Figure 1. Example of training data.

Our dataset consists of disease outbreak alerts retrieved in 2007 by the HealthMap system. We tagged them with the part-of-speech tagger provided by NEC's project SENNA (Collobert & Weston, 2007), which has a reported accuracy of 96.85%. Provided that, in English, location names often begin with capital letters or appear as acronyms, we assigned to the words in the alerts, in addition to their part-of-speech tags, a capitalization status, *ie* none, first character, upper case. We used the rule-based approach to tag the words in the alert that match geographical references found in the gazetteer. Since the alerts in the dataset have been displayed (with the supervision of an expert) in the HealthMap world map, we were able in addition, to distinguish among the assigned tags, the ones that actually referred to a location where the event was taking place (location IN or *locIN*), from those that were foreign mentions (location OUT or *locOUT*).

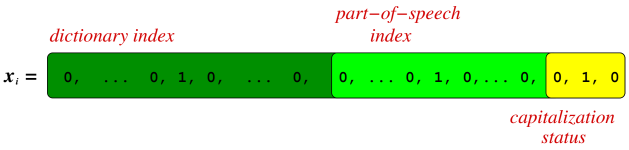

Figure 2. Words sparse representation.

From a machine learning perspective, our dataset is composed of alerts examples $\mathbf{x} = [x_1, \ldots, x_L]$ of length L, and their corresponding location labels $\mathbf{y} = [y_1, \ldots, y_L]$, $y_i \in \{None, locIN, locOUT\}$. The words x_i are represented by their part-of-speech tag, their capitalization status and occasionally by their index in the dictionary $\mathcal{D}$, extracted from the training dataset. Figure 2 illustrates the vectorial representation of words. The $|\mathcal{D}|$ (size of $\mathcal{D}$) first components of x_i correspond to the dictionary indexes and are all equal to zero, except for the position coinciding with the word index in $\mathcal{D}$. Similarly, the next K features of x_i correspond to the part-of-speech tag indexes in the K part-of-speech tag list. And finally, the last three features stand for the three possible capitalization status. As explained previously, one important characteristic of this experiment, is the fact that words are only partially accessible to the learning algorithm. We applied a lower bound $minfreq$ on the word frequency in the dataset to decide whether or not a word index was to be hidden. Only frequent word indexes are visible to the model. This was implemented as a dictionary cut-off, in which infrequent words are removed from the dictionary. Table 1 reports the resulting sub-dictionaries size for varying $minfreq$, in two training datasets T_0 and T_1 which respectively have 1000 and 2500 alerts. An out-of-dictionary word will have its $|\mathcal{D}|$ first components equal to zero.

$minfreq$	0	4	10	20
T_0 dict. size	$15,000$	$5,000$	$3,000$	$1,700$
T_1 dict. size	$25,300$	$9,500$	$5,900$	$3,900$

Table 1. Sizes of (sub-)dictionaries extracted from the training datasets T_0 (1000 alerts) and T_1 (2500 alerts).

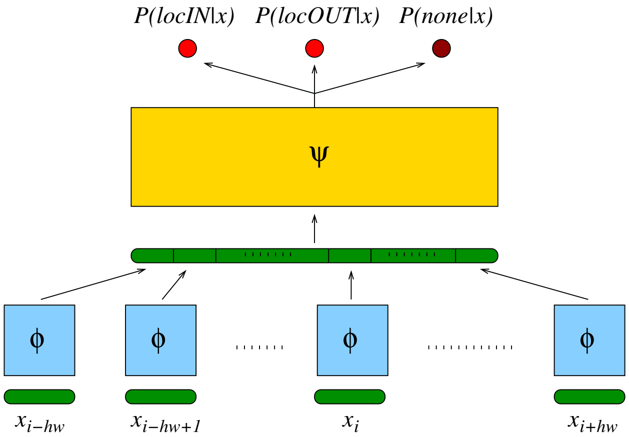

Figure 3. Illustration of the neural network.

We trained a neural network, by negative log-likelihood minimization to output a probability estimate of the label y_i value corresponding to the word x_i in i^{th} position of an alert $\mathbf{x}$,

$$NN(i, \mathbf{x}) = P(y_i | x_{i-hw}, \ldots, x_i, \ldots, x_{i+hw})$$

given a window ($n - 1 = 2 \times hw$) of preceding and following words.

$minfreq$	0	4	10	20
% w/o index (T_0 dict.)	4.9	9.7	13.6	18.1
% loc w/o index (T_0 dict.)	16.8	36.6	47.3	61.4
% w/o index (T_1 dict.)	3.0	5.8	8.1	10.6
% loc w/o index (T_1 dict.)	6.6	22.1	31.3	38.7

Table 2. Percentage among the validation set words and words labeled as locations, of words without index in the dictionary.

The neural network, illustrated in Figure 3, can be decomposed as follow. First, each word in the window sequence is given in input to the same multi-layer perceptron (MLP) which has been replicated $n = 2 \times hw + 1$ times, in a siamese network fashion (Bromley et al., 1993). This first MLP can be seen as a function ϕ mapping the extremely sparse representation x_i of the words into a new representation, $\phi(x_i) \in \mathbf{R}^d$, which has the advantage of being learned during the training. This approach was applied with success for language modelling in (Bengio et al., 2003) and more recently for semantic role parsing in (Collobert & Weston, 2007). The outputs $\phi(x_{i-hw}), \ldots, \phi(x_i), \ldots, \phi(x_{i+hw})$ are concatenated into a vector $z \in \mathbf{R}^{d \times n}$ which is itself given in input to a second multi-layer perceptron. This second MLP, called ψ in Figure 3, has as output layer a *softmax* filtering function which allows us to consider the outputs of the neural network as probabilities.

3. Results

We trained several such neural networks on the two datasets T_0 (1,000 alerts) and T_1 (2,500 alerts), with extracted dictionaries of varying sizes according to our lower bound $minfreq$, as described in Table 1. We tested the models obtained on a separated validation set of 1000 alerts (465,297 words to tag, 6,156 with target *locIN* and 5,013 with *locOUT*). Table 2 summarizes the percentage in that set, of words that, as a consequence of the dictionaries cut-off, do not have an index in the dictionaries extracted from T_0 and T_1. The 2nd and 4th lines of Table 2 show the out-of-dictionary percents among the words that were assigned a location tag. The much higher proportions confirms the intuition that individual location mentions are infrequent words. Note that in the first column of Table 2, where no word is removed from the original dictionaries, there is still a certain amount of out-of-dictionary words. This is due to the fact that the vocabulary of the validation dataset is not completely covered by the dictionaries extracted from the training datasets.

Given the approximate nature of the solution found when training neural networks by stochastic gradient descent we repeated the learning process for each condition 5 times to estimate the variance. Figure 4 displays the results in terms of the obtained F_1 score for the parsing of the words targeted with tags location IN and location OUT, as well as the F_1 score obtained if we do not make the distinction between the two of them (referred as *loc*), with models trained both on T_0 and T_1. There is a discrepancy in the results for the tag *locIN / locOUT* and the reported *loc* tag, showing that the neural networks confuse *locINs* for *locOUTs* targets, and *vice versa*. That seems to suggest that our set of features is not suited to fully make this distinction. The size of the window in particular which we kept relatively small ($hw = 4$) for these preliminary experiments, may help to narrow this gap. If we consider the performance on the whole location tags, however, the results seem to be encouraging. Another encouraging facts is that increasing the size of the training dataset improves the performance of the models.

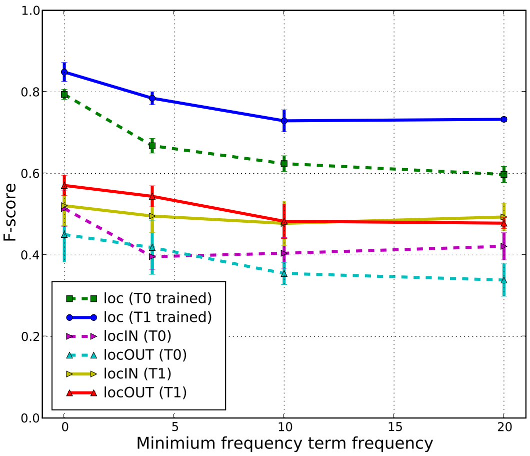

Figure 4. F1-score of *locIN*, *locOUT*, and the grouped location labels for two sizes of training dataset (train0: 1000 alerts, train1: 2500 alerts)

In Figure 5, to emphasize the fact that this approach do differ from the gazetteer-based approach, the results obtained with models trained on T_1 have been sliced into the F_1 scores of words that were represented with and without their dictionary index feature. Results for the words that were not targeted as locations (None tags), are also reported in Figure 5, they oscillated around an F_1 score of 99% for words with their index feature and 95% for those without.

The observed increase in performance for no-index words proportionally to the size of the dictionary cut-off suggest a potential for discovering phrases out of the initial gazetteer, and that, without having a too high lost in performance for unhidden words. A visual inspection of the false positive among the words without index, reveals that indeed many among the words labeled as *None* that the system decided

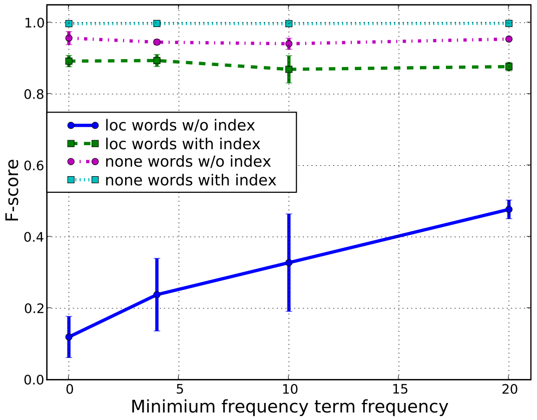

Figure 5. F1-score of *loc* and *none* targets for words with and without dictionary index feature in their representation.

were location are in fact location references that are out of the gazetteer.

4. Conclusion

We have presented a promising approach to incorporate the prior knowledge encoded in a rule-based procedure into a more general statistical framework. We have demonstrated that the described model has the ability to discover geographic references based solely on the context they appear in. The experiments also attested that providing additional training material improves the performance of the model, suggesting that despite the fabricated nature of the data it is still able to dispense interesting information. We plan in the close future to try more complex features for the word representations such as *eg* their semantic role labels. This technique could be integrated to a more conventional method for geo-parsing based on a geographically annotated dataset. It would be interesting to evaluate the contribution this approach could provide to the final task of indexing disease outbreak reports for geographic information retrieval.

References

Bengio, Y., Ducharme, R., Vincent, P., & Gauvin, C. (2003). A Neural Probabilistic Language Model. *JMLR*, *3*, 1137–1155.

Bromley, J., Guyon, I., LeCun, Y., Sackinger, E., & Shah, R. (1993). Signature Verification using a Siamese Time Delay Neural Network. *Advances in Neural Information Processing Systems 6*.

Brownstein, J. S., & Freifeld, C. C. (2007). Healthmap: the development of automated real-time internet surveillance for epidemic intelligence. *Euro Surveill*, *12(48)*, 3322.

Carreras, X., & Màrquez, L. (2005). Introduction to the CoNLL-2005 Shared Task: Semantic Role Labeling. *Proceedings of the 9th Conference on Natural Language Learning, CoNLL-2005*. Ann Arbor, MI USA.

Collobert, R., & Weston, J. (2007). Fast semantic extraction using a novel neural network architecture. *Proceedings of 45th Annual Meeting of the Association for Computational Linguistics*.

Freifeld, C. C., Mandl, K. D., Reis, B. Y., & Brownstein, J. S. (2008). Healthmap: Global infectious disease monitoring through automated classification and visualization of internet media reports. *J Am Med Inform Assoc*, *15(2)*, 150–157.

Tjong Kim Sang, E. F., & De Meulder, F. (2003). Introduction to the conll-2003 shared task: Language-independent named entity recognition. *Proceedings of the Seventh Conference on Natural Language Learning (CoNLL-03)* (pp. 142–147). Edmonton, Canada.

Woodruff, A. G., & Plaunt, C. (1994). Gipsy: Automated geographic indexing of text documents. *Journal of the American Society for Information Science*, *45:9*, 645–655.