Improving Text representations through Probabilistic Integration of Synonymy Relations

Romaric Besançon, Jean-Cédric Chappelier, Martin Rajman, and Antoine Rozenknop

Artificial Intelligence Laboratory, Computer Science Department
Swiss Federal Institute of Technology, CH-1015 Lausanne
{Romaric.Besancon,Martin.Rajman,Jean-Cedric.Chappelier,
Antoine.Rozenknop}@epfl.ch

Abstract. The present contribution focuses on the integration of word senses in a vector representation of texts, using a probabilistic model. The vector representation under consideration is the DSIR model, that extends the standard Vector Space (VS) model by taking both occurrences and co-occurrences of words into account. Integration of word senses into the co-occurrence model is done using a Markov Random Field model with hidden variables, using semantic information derived from synonymy relations extracted from a synonym dictionary.

1 Introduction

Information Retrieval (IR) systems are designed to compute the similarities between a query and a collection of documents; standard Clustering techniques rely on a similarity measure that allows to build clusters by grouping similar documents.

These text similarity based systems generally use a vector space (VS) representation of the documents (Salton and Buckley (1988)) to derive a mathematical similarity measure in the considered vector space (for example, the cosine of the angle between the vectors representing the documents). The dimensions of the vector space are usually associated with specific linguistic units, that can be words, stems or lemmas, called *indexing terms*. The idea presented in this paper is to enhance the accuracy of the text representations by specifying the senses of the polysemous terms. To do so, a word sense disambiguation phase is integrated in the representation process. The semantic information required for this process is extracted from the synonymy relations contained in a synonym dictionary.

In section 2, we briefly review the DSIR (*Distributional Semantics for Information Retrieval*) model for text representation. In section 3, we present the way synonymy relations are considered, and how we derive a notion of *concept* from a set of synonymy relations. Section 4 deals with the integration of the Word Sense Disambiguation in the probabilistic model, and the use of an EM algorithm to estimate the model parameters. Finally, in section 5, we present the experimental framework that is foreseen for the evaluation of our approach.

2 The DSIR representation model

Standard VS models represent the documents by their *lexical profiles*[1], only taking the occurrences of a term in a document into account. The idea of the DSIR model is to take both occurrences and *co-occurrences* of terms into account in the vector representation of documents (see Rungsawang and Rajman (1995), Rajman et al. (2000)).

The original idea of Distributional Semantics states that the semantics of a word is related to the set of contexts in which that word appears (Rajman and Bonnet (1992)). In the DSIR approach, the contexts are included through co-occurrence relations. A linguistic unit u_j is represented by its *co-occurrence profile* $(c_{j1}, \ldots, c_{jM})$, where M is the size of the indexing set and c_{ji} is the co-occurrence frequency of u_j with the indexing term t_i. The documents are then represented by the weighted average of the co-occurrence profiles of the words they contain, *i.e.* $d = (d_1, \ldots, d_M)$, where $d_i = \sum_{u_j \in U} f_j \, c_{ji}$, U denoting the set of all linguistic units, and f_j the frequency of u_j in the document d. As shown in Besancon et al. (1999), Rajman et al. (2000), the DSIR model can also have a probabilistic interpretation.

The simplest way to compute co-occurrence frequencies is to consider co-occurrences relations between all linguistic units in a sentence or a fixed-length window. A more efficient way is to use some additional syntactic information to produce syntactic groups along with their heads, and to only consider co-occurrences within a syntactic group or between heads of syntactic groups, thus avoiding considering spurious co-occurrences between dependences from different groups (Besançon et al. (1999)). The sentence is then rewritten as a *co-occurrence graph* in which the nodes are associated with the considered linguistic units and the arcs represent the co-occurrence relations.

For instance, let us consider the following sentence: "*It is the quiet pigs that eat the meal.*" (Irish proverb). After a pre-processing step that only keeps the lemmas of content bearing words[2], co-occurrence information can be computed. The resulting graphs, depending on whether syntactic filtering is or is not taken into account, are presented in Figure 1.

quiet pig eat meal (a)

*(quiet *pig)(*eat)(*meal)* (b)

Fig. 1. Examples of graph of co-occurrences , graph (a) without syntactic filtering, graph (b) using syntactic filtering (the stars indicate the heads of the syntactic groups).

[1] the lexical profile is a vector the components of which represent the importance of the indexing terms in the document.

[2] for instance, one can decide to keep nouns, verbs and adjectives only.

3 The synonym model

A standard synonym dictionary provides a list of synonyms for each sense of its word entries. For example, the senses and the corresponding synonyms for the word *eat* (as given by the Merriam-Webster Thesaurus) are:

- Sense 1 ⇒ *consume, devour, feed (on), ingest, partake (of), take*;
- Sense 2 ⇒ *consume, devour, eat up, exhaust, use up*;
- Sense 3 ⇒ *bite, corrode, eat away, erode, gnaw, scour, wear (away)*;

We denote $w_i^1, \ldots, w_i^{n_i}$ the n_i senses of a word w_i, and $syn(w_i^j) \subset U$ the set of synonyms[3] for the sense j of the word w_i.

Let us consider the *is-synonym-to* relation[4]. As standard dictionaries are usually not complete with respect to this relation (transitivity is often not verified), we first build the transitive closure of the relation in order to extract the subgraphs corresponding to connected components of senses. By construction, a connected component is then a set of word senses expressing some shared meaning (as expressed by the synonymy relation). This meaning bearing set of word senses will be called a *concept*.

In the previous example, the word *consume* is a synonym for two senses of *eat*. We must therefore assume that this word has itself at least two different senses. More generally, the words appearing in the raw synonym lists provided by the dictionary are often polysemous and must be disambiguated to allow the construction of concepts. Heuristic approaches have to be used to associate a sense with each of the words appearing in the synonym lists. Our heuristic is to associate to each word $w_i \in syn(w_k^l)$ the sense w_i^j for which the intersection between the synonym list representing w_i^j and the synonym list representing w_k^l is maximal. More precisely, the associated sense j^* is defined by:

$$j^* = \underset{j}{\operatorname{argmax}} \left| syn(w_i^j) \bigcap \{w_k^l \cup syn(w_k^l) \setminus w_i^j\} \right|$$

4 Integrating word sense disambiguation in the DSIR representation

The objective of the integration of word senses in the DSIR model is to compute co-occurrences between concepts instead of co-occurrences between linguistic units in order to build richer semantic representations with the co-occurrence profiles.

The main problem to do so is to choose the right concepts to be associated with the words in the corpus to be processed. The approach chosen for such a disambiguation is to follow the intuition underlying Distributional

[3] U is the set of word entries in the dictionary, and corresponds here to the set of linguistic units considered in section 2.

[4] we assume this relation to be symmetric, which corresponds to the intuition.

Semantics: we make the hypothesis that the sequence of concepts to be associated with a given sequence of words is the one that maximises the probability of assigning concepts to the nodes of the co-occurrence graph derived from the word sequence. In other words, for any given word assignment w to a co-occurrence graph g, the selected concept assignment c^* is such that $c^* = \text{argmax}_c\, p(c|w,g) = \text{argmax}_c\, p(w|c,g)\; p(c|g)$.

To further compute $p(w|c,g)\; p(c|g)$, we make the following hypotheses:

- **H1:** the conceptual conditioning imposed on any given word assignment is limited to the associated concept: $p(w|c,g) = \prod_i p(w_i|c_i)$;
- **H2:** the conceptual conditioning imposed on any given concept assignment is limited to its neighbourhood in the co-occurrence graph: $p(c_i|g) = p(c_i|\mathcal{V}_i)$, where $\mathcal{V}_i$ is the neighbourhood of node i.

The hypothesis **(H2)** entails a *Markov Random Field* (MRF) structure for the model and, according to the theorem of Hammersley (Besag (1974)), imposes that the probability distribution on concept configurations $p(c|g)$ is a Gibbs Distribution, *i.e.* a distribution of the form:

$$p(c|g) = \frac{1}{Z_0} \exp \sum_{\gamma \in \Gamma} V_\gamma(c)$$

where Z_0 is a normalisation factor, Γ is a set of cliques in g, and $V_\gamma(c)$ is a function associated to these cliques, called the *potential*. In our case, since we are only considering co-occurrence relations, the cliques to be considered will be of size 2. The probability function to maximise to obtain the optimal concept assignment c^* is then:

$$p(w|c,g)\; p(c|g) = \frac{1}{Z} \exp \left[\sum_{i \in g} V_e(w_i, c_i) + \sum_{(i,j) \in g} V(c_i, c_j) \right]$$

where $V_e(w_i, c_i)$ is the potential associated to the emission probability $p(w_i|c_i)$. This is also a Gibbs distribution that can be generically written (without formally distinguishing $V(c_i, c_j)$ and $V_e(w_i, c_i)$): $p(w, c|g) = \frac{1}{Z} \exp \sum_{i,j} V(x_i, x_j)$. where $x_i, x_j \in U \cup \mathcal{C}$, $\mathcal{C}$ denoting the set of concepts.

This probability distribution corresponds to a parametric model that will be used for the disambiguation task. In order to assign relevant values to the parameters $V(x_i, x_j)$, we use the standard approach consisting in choosing the values that maximise the log-likelihood of a training corpus C.

To do so, we use an iterative algorithm based on the *Estimation-Maximisation (EM) algorithm* (Dempster et al. (1977)) and the *Improved Iterative Scaling (IIS) algorithm* (Lafferty (1996)), that try to iteratively find the parameters set $\theta = \{V(x_i, x_j)\}$ that maximises the expectation of the log-likelihood $L_\theta(C) = \sum_{w \in C} \log p_\theta(w)$. This kind of technique is often used in image reconstruction (Chalmond (1986)).

Due to the global character of the normalisation constant Z, the computation of the likelihood is not tractable: the usual way of circumventing this problem is to use the pseudo-likelihood function $\mathrm{PL}_\theta(x|g) = \sum_{i \in g} \log p_\theta(x_i|\mathcal{V}_i)$ (Besag (1974), Chalmond (1986)). In our case, we can show that maximising the function $A(\theta, \theta')$ given in Equation (1) (provided that $A(\theta, \theta') > 0$) will assure the parameters θ' to give a better model than θ, according to the maximum pseudo-likelihood estimator:

$$A(\theta,\theta') = \sum_{(w,g)} \sum_{c} p_\theta(c|w,g) \sum_{i \in g} \left[\sum_{j \in \mathcal{V}_i} \Delta V(x_i,x_j) - \sum_{x \in U \cup \mathcal{C}} p_\theta(x|\mathcal{V}_i) \frac{1}{|\mathcal{V}_i|} \sum_{j \in \mathcal{V}_i} \exp |\mathcal{V}_i| \, \Delta V(x,x_j) \right] \tag{1}$$

where $\Delta V(x_i,x_j) = V_{\theta'}(x_i,x_j) - V_\theta(x_i,x_j)$.

As the derivatives $\frac{\partial A(\theta,\theta')}{\partial V(x_i,x_j)}$ only depend on $V(x_i,x_j)$, we can therefore find the parameters of our model by iteratively computing the values $\Delta V(x_i,x_j)$ that maximise $A(\theta,\theta')$, and replacing $V(x_i,x_j)$ by $V(x_i,x_j) + \Delta V(x_i,x_j)$ until convergence.

5 Evaluation

The approach presented in this paper is validated on an IR task, using a corpus of newspaper articles from the French newspaper *Le Monde*, from the AMARYLLIS evaluation campaign for IR systems. Notice that the validation does not directly concentrate on the results obtained for word sense disambiguation, but rather on the indirect impact of the use of the produced concepts for a more general IR task. Indeed, for this task, the texts (documents and queries) are represented in a vector space with the DSIR model based on co-occurrences of senses instead of co-occurrences of words.

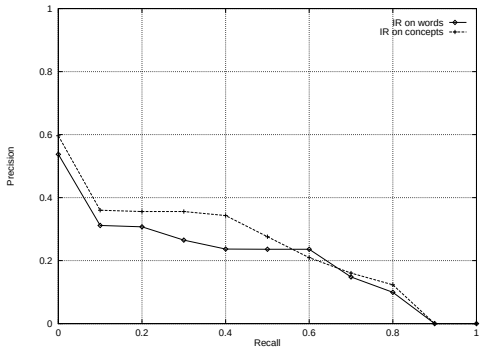

Fig. 2. Results of the word sense integration in an Information Retrieval task on the "Le Monde" corpus.

The disambiguation work on the synonym lists derived from a synonym dictionary has been carried out separately and a set of concepts have been derived (Pfister (2000)). In the IR task, a lexicon of 7073 words is used, corresponding to a total of 2936 concepts (the average number of words per

concepts is 2.7). The reference corpus is composed of 9574 documents, and 4 queries. The precision/recall results, presented in Figure 2 show a slight improvement of the performance using the concept representation of the documents and the queries. Further validation experiments are currently ongoing to provide more detailed insights into the characteristics of the method that explain the observed performance improvements.

6 Conclusion

In this paper, we present a model that uses Markov Random Field (MRF) to represent the co-occurrences, in order to integrate word senses in a probabilistic model for text representation. MRF models seem to provide a good framework for the representation of this kind of undirected neighbourhood information, and might be also considered for the direct modelling of word co-occurrence in the representation of documents. The first results for the evaluation of this model on a IR task are promising, even though a deeper evalutation should be conducted. In addition, having a co-occurrence model for senses is also useful for Word Sense Disambiguation *per se*, but should be evaluated in a specific manner.

References

BESAG, J. (1974): Spatial interaction and the statistical analysis of lattice systems. *Journal of Royal Statistics Society*, 36:192–236.

BESANÇON, R., RAJMAN, M., and CHAPPELIER, J.-C. (1999): Textual similarities based on a distributional approach. In *International Workshop on Similarity Search (IWOSS99)*, Florence, Italy.

CHALMOND, B. (1986): An Iterative Gibbsian Technique for Reconstruction of m-ary Images. *Pattern Recognition*, 22(6):747–761.

DEMPSTER, A., LAIRD, N., and RUBIN, D. (1977): Maximum likelihood from incomplete data via the EM algorithm. *Jounal of Royal Statistics Society*, 39:185–197.

LAFFERTY, J. (1996): Gibbs-markov models. *Computing Science and Statistics*, 27:370–377.

PFISTER, J.-P. (2000). Désambiguisation d'un dictionnaire de synonymes. Technical report, EPFL.

RAJMAN, M., BESANÇON, R., and CHAPPELIER, J.-C. (2000): Le modèle DSIR : Une approche à base de sémantique distributionnelle pour la recherche documentaire. *Traitement Automatique des Langues*, 41(2).

RAJMAN, M. and BONNET, A. (1992): Corpora-base linguistics: new tools for natural language processing. In *1st Annual Conference of the Association for Global Strategic Information*, Bad Kreuznach, Germany.

RUNGSAWANG, A. and RAJMAN, M. (1995). Textual information retrieval based on the concept of distributional semantics. In *proc. of JADT'95 (3rd International Conference on Statistical Analysis of Textual Data)*, Rome.

SALTON, G. and BUCKLEY, C. (1988): Term weighting approaches in automatic text retrieval. *Information Processing and Management*, 24:513–523.