Bielefeld SC: Orthonormal Topic Modelling for Grammar Induction

John P. McCrae
CITEC, Bielefeld University
Inspiration 1
Bielefeld, Germany
jmccrae@cit-ec.uni-bielefeld.de

Philipp Cimiano
CITEC, Bielefeld University
Inspiration 1
Bielefeld, Germany
cimiano@cit-ec.uni-bielefeld.de

Abstract

In this paper, we consider the application of topic modelling to the task of inducting grammar rules. In particular, we look at the use of a recently developed method called orthonormal explicit topic analysis, which combines explicit and latent models of semantics. Although, it remains unclear how topic model may be applied to the case of grammar induction, we show that it is not impossible and that this may allow the capture of subtle semantic distinctions that are not captured by other methods.

1 Introduction

Grammar induction is the task of inducing high-level rules for application of grammars in spoken dialogue systems. In practice, we can extract relevant rules and the task of grammar induction reduces to finding similar rules between two strings. As these strings are not necessarily similar in surface form, what we really wish to calculate is the semantic similarity between these strings. As such, we could think of applying a semantic analysis method. As such we attempt to apply topic modelling, that is methods such as Latent Dirichlet Allocation (Blei et al., 2003), Latent Semantic Analysis (Deerwester et al., 1990) or Explicit Semantic Analysis (Gabrilovich and Markovitch, 2007). In particular we build on the recent work to unify latent and explicit methods by means of orthonormal explicit topics.

In topic modelling the key choice is the document space that will act as the corpus and hence topic space. The standard choice is to regard all articles from a background document collection – Wikipedia articles are a typical choice – as the topic space. However, it is crucial to ensure that these topics cover the semantic space evenly and completely. Following McCrae et al. (McCrae et al., 2013) we remap the semantic space defined by the topics in such a manner that it is orthonormal. In this way, each document is mapped to a topic that is distinct from all other topics.

The structure of the paper is as follows: we describe our method in three parts, first the method in section 2, followed by approximation method in section 3, the normalization methods in section 4 and finally the application to grammar induction in section 5, we finish with some conclusions in section 6.

2 Orthonormal explicit topic analysis

ONETA (McCrae et al., 2013, Orthonormal explicit topic analysis) follows Explicit Semantic Analysis in the sense that it assumes the availability of a background document collection $B = \{b_1, b_2, ..., b_N\}$ consisting of textual representations. The mapping into the explicit topic space is defined by a language-specific function Φ that maps documents into $\mathbb{R}^N$ such that the j^{th} value in the vector is given by some *association measure* $\phi_j(d)$ for each background document b_j. Typical choices for this association measure ϕ are the sum of the TF-IDF scores or an information retrieval relevance scoring function such as BM-25 (Sorg and Cimiano, 2010).

For the case of TF-IDF, the value of the j-th element of the topic vector is given by:

$$\phi_j(d) = \overrightarrow{\text{tf-idf}}(b_j)^{\text{T}}\ \overrightarrow{\text{tf-idf}}(d)$$

Thus, the mapping function can be represented as the product of a TF-IDF vector of document d multiplied by a word-by-document ($W \times N$) TF-IDF matrix, which we denote as a $\mathbf{X}$:[1]

[1] $^{\text{T}}$ denotes the matrix transpose as usual

Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pages 119–122,
Dublin, Ireland, August 23-24, 2014.

$$\Phi(d) = \begin{pmatrix} \overrightarrow{\text{tf-idf}}(b_1)^{\mathrm{T}} \\ \vdots \\ \overrightarrow{\text{tf-idf}}(b_N)^{\mathrm{T}} \end{pmatrix} \overrightarrow{\text{tf-idf}}(d) = \mathbf{X}^{\mathrm{T}} \cdot \overrightarrow{\text{tf-idf}}(d)$$

For simplicity, we shall assume from this point on that all vectors are already converted to a TF-IDF or similar numeric vector form.

In order to compute the similarity between two documents d_i and d_j, typically the cosine-function (or the normalized dot product) between the vectors $\Phi(d_i)$ and $\Phi(d_j)$ is computed as follows:

$$\text{sim}(d_i, d_j) = \cos(\Phi(d_i), \Phi(d_j)) = \frac{\Phi(d_i)^{\mathrm{T}}\Phi(d_j)}{||\Phi(d_i)||||\Phi(d_j)||}$$

$$\text{sim}(d_i, d_j) = \cos(\mathbf{X}^{\mathrm{T}}d_i, \mathbf{X}^{\mathrm{T}}d_j) = \frac{d_i^{\mathrm{T}}\mathbf{X}\mathbf{X}^{\mathrm{T}}d_j}{||\mathbf{X}^{\mathrm{T}}d_i||||\mathbf{X}^{\mathrm{T}}d_j||}$$

The key challenge with topic modelling is choosing a good background document collection $B = \{b_1, ..., b_N\}$. A simple minimal criterion for a good background document collection is that each document in this collection should be maximally similar to itself and less similar to any other document:

$$\forall i \neq j \;\; 1 = \text{sim}(b_j, b_j) > \text{sim}(b_i, b_j) \geq 0$$

As shown in McCrae et al. (2013), this property is satisfied by the following projection:

$$\Phi_{\text{ONETA}}(d) = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}d$$

And hence the similarity between two documents can be calculated as:

$$\text{sim}(d_i, d_j) = \cos(\Phi_{\text{ONETA}}(d_i), \Phi_{\text{ONETA}}(d_j))$$

3 Approximations

ONETA relies on the computation of a matrix inverse, which has a complexity that, using current practical algorithms, is approximately cubic and as such the time spent calculating the inverse can grow very quickly.

We notice that $\mathbf{X}$ is typically very sparse and moreover some rows of $\mathbf{X}$ have significantly fewer non-zeroes than others (these rows are for terms with low frequency). Thus, if we take the first N_1 columns (documents) in $\mathbf{X}$, it is possible to rearrange the rows of $\mathbf{X}$ with the result that there is some W_1 such that rows with index greater than W_1 have only zeroes in the columns up to N_1. In other words, we take a subset of N_1 documents and enumerate the words in such a way that the terms occurring in the first N_1 documents are enumerated $1, \ldots, W_1$. Let $N_2 = N - N_1$, $W_2 = W - W_1$. The result of this row permutation does not affect the value of $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ and we can write the matrix $\mathbf{X}$ as:

$$\mathbf{X} = \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{0} & \mathbf{C} \end{pmatrix}$$

where $\mathbf{A}$ is a $W_1 \times N_1$ matrix representing term frequencies in the first N_1 documents, $\mathbf{B}$ is a $W_1 \times N_2$ matrix containing term frequencies in the remaining documents for terms that are also found in the first N_1 documents, and $\mathbf{C}$ is a $W_2 \times N_2$ containing the frequency of all terms not found in the first N_1 documents.

Application of the well-known divide-and-conquer formula (Bernstein, 2005, p. 159) for matrix inversion yields the following easily verifiable matrix identity, given that we can find $\mathbf{C}'$ such that $\mathbf{C}'\mathbf{C} = \mathbf{I}$.

$$\begin{pmatrix} (\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-1}\mathbf{A}^{\mathrm{T}} & -(\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{B}\mathbf{C}' \\ \mathbf{0} & \mathbf{C}' \end{pmatrix} \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{0} & \mathbf{C} \end{pmatrix} = \mathbf{I} \quad (1)$$

The inverse $\mathbf{C}'$ is approximated by the *Jacobi Preconditioner*, $\mathbf{J}$, of $\mathbf{C}^{\mathrm{T}}\mathbf{C}$:

$$\begin{aligned} \mathbf{C}' &\simeq \mathbf{J}\mathbf{C}^{\mathrm{T}} \quad (2) \\ &= \begin{pmatrix} ||c_1||^{-2} & & 0 \\ & \ddots & \\ 0 & & ||c_{N_2}||^{-2} \end{pmatrix} \mathbf{C}^{\mathrm{T}} \end{aligned}$$

4 Normalization

A key factor in the effectiveness of topic-based methods is the appropriate normalization of the elements of the document matrix $\mathbf{X}$. This is even more relevant for orthonormal topics as the matrix inversion procedure can be very sensitive to small changes in the matrix. In this context, we consider two forms of normalization, term and document normalization, which can also be considered as row/column normalizations of $\mathbf{X}$.

A straightforward approach to normalization is to normalize each column of $\mathbf{X}$ to obtain a matrix as follows:

$$\mathbf{X}' = \left(\frac{x_1}{||x_1||} \cdots \frac{x_N}{||x_N||} \right)$$

If we calculate ${\mathbf{X}'}^{\mathrm{T}}\mathbf{X}' = \mathbf{Y}$ then we get that the (i, j)-th element of $\mathbf{Y}$ is:

$$y_{ij} = \frac{x_i^{\mathrm{T}} x_j}{||x_i||||x_j||}$$

Thus, the diagonal of $\mathbf{Y}$ consists of ones only and due to the Cauchy-Schwarz inequality we have that $|y_{ij}| \leq 1$, with the result that the matrix $\mathbf{Y}$ is already close to $\mathbf{I}$. Formally, we can use this to state a bound on $||{\mathbf{X}'}^{\mathrm{T}}\mathbf{X}' - \mathbf{I}||_{\mathcal{F}}$, but in practice it means that the orthonormalizing matrix has more small or zero values. Previous experiments have indicated that in general term normalization such as TF-IDF is not as effective as using the direct term frequency in ONETA, so we do not apply term normalization.

5 Application to grammar induction

The application to grammar induction is simply carried out by taking the rules and creating a single ground instance. That is if we have a rule of the form

`LEAVING FROM <CITY>`

We would replace the instance of `<CITY>` with a known terminal for this rule, e.g.,

leaving from Berlin

This reduces the task to that of string similarity which can be processed by means of any string similarity function, for example such as the ONETA function described above. As such the procedure is as follows:

1. Ground the input grammar rule to an English string d
2. Ground each candidate matching rule to an English string d_i
3. Calculate for each d_i, the similarity $\mathrm{sim}_{\mathrm{ONETA}}(d, d_i)$
4. Add the rule to the grammar class with the highest similarity

This approach has the obvious drawback that it removes all information about the valence of the rule, however the effect of this loss of information remains unclear.

For application, we used 20,000 Wikipedia articles, filtered to contain only those of over 100 words, giving us a corpus of 15.6 million tokens. We applied ONETA using document normalization but no term normalization and the value $N_1 = 5000$. These parameters were chosen based on the best results in previous experiments.

6 Conclusions

The results show that such a naive approach is not directly applicable to the case of grammar induction, however we believe that it is possible that the subtle semantic similarities captured by topic modelling may yet prove useful for grammar induction. However it is clear from the presented results that the use of a topic model alone does not suffice to solve this task. We notice that from the data many of the distinctions rely on antonyms and stop words, especially distinctions such as 'to'/'from', which are not captured by a topic model as topic models generally ignore stop words, and generally consider antonyms to be in the same topic, as they frequently occur together in text. The question of when semantic similarity such as provided by topic modelling is applicable remains an open question.

References

Dennis S Bernstein. 2005. *Matrix mathematics, 2nd Edition*. Princeton University Press Princeton.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022.

Scott C. Deerwester, Susan T Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. 1990. Indexing by latent semantic analysis. *JASIS*, 41(6):391–407.

Evgeniy Gabrilovich and Shaul Markovitch. 2007. Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, volume 6, page 12.

John P. McCrae, Philipp Cimiano, and Roman Klinger. 2013. Orthonormal explicit topic analysis for cross-lingual document matching. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1732–1740.

Philipp Sorg and Philipp Cimiano. 2010. An experimental comparison of explicit semantic analysis implementations for cross-language retrieval. In *Natural Language Processing and Information Systems*, pages 36–48. Springer.