TIAD Shared Task 2019: Orthonormal Explicit Topic Analysis for Translation Inference across Dictionaries

John P. McCrae

Insight Centre for Data Analytics/Data Science Institute
National University of Ireland, Galway
john@mccr.ae

Abstract. The task of inferring translations can be achieved by the means of comparable corpora and in this paper we apply explicit topic modelling over comparable corpora to the task of inferring translation candidates. In particular, we use the Orthonormal Explicit Topic Analysis (ONETA) model, which has been shown to be the state-of-the-art explicit topic model through its elimination of correlations between topics. The method proves highly effective at selecting translations with high precision.

Keywords: Topic Modelling · Explicit Topics · Translation Inference.

1 Introduction

Explicit topic modelling, such as proposed by the Explicit Semantic Analysis (ESA) [3] method, is a method that in contrast to latent topic modelling, such as Latent Dirichlet Allocation (LDA) [2], or word embeddings relies on the user to explicitly give a list of topics. These topics are a set of documents that are supposed to correspond to the major topical areas of the domain, however in most works, including this one, a set of Wikipedia articles are chosen as the explicit topics. This method while obviously requiring more manual effort than latent methods, does provide a number of advantages, most notably that the topics can easily be aligned across languages and this has been implemented by Cross-lingual Explicit Semantic Analysis (CL-ESA) [7]. In contrast, latent methods require a complex and error-prone step of aligning the latent topics across languages [8]. One of the principle criticisms of explicit semantic analysis is that the choice of underlying implementation can strongly affect the quality of the resulting system [6]. One of the main reasons for this is the fact that the topics chosen for the explicit analysis are often highly similar and that this causes a lack of orthogonality between the topics [4, 1]. For this reason we use the Orthonormal Explicit Topic Analysis (ONETA) [4] method in order to find cross-lingual equivalents between terms.

Translation inference is the task of inferring a translation equivalent between two languages by means of other bilingual dictionaries in other language pairs.

The principle issue is that the translation graph is not transitive, so by following a translation pair from English to Spanish, and then a translation pair from Spanish to French and incorrect translation may be inferred if there are multiple senses of the Spanish word that is used as a pivot. However, previous TIAD tasks [5] have shown that this is a moderately high precision method. For this edition of the task, we proposed filtering the results of pivot translations by means of inferred cross-lingual similarity using ONETA, with the idea that translations that are both found by the pivot and ranked as highly-similar by the ONETA method are likely to be high quality translations. In this way, we provide a method that allows the lexicographer to easily adjust the method to a level of precision that is most suitable for validating translation candidates generated by pivot-based translation.

2 Orthonormal Explicit Topic Analysis

Orthonormal explicit topic analysis follows from explicit semantic analysis by assuming there is a background collection of documents we call $B = \{b_1, \dots, b_n\}$, and in the cross lingual setting it is assumed that there is a paired set of documents $B' = \{b'_1, \dots, b'_n\}$, with each document being paired with a similar document in a second language. This is most frequently achieved by using Wikipedia, where interlingual links link two articles in different languages. It is assumed that we have some language-specific function Φ that maps a document to a vector in $\mathbb{R}^n$, such that the j^{th} element of the vector is an association with the $\phi_j(d)$ with the document b_j. In our method, this vector is given by a metric such as TF-IDF such that in our score is:

$$\phi_j(d) = \overrightarrow{\text{tf-idf}(b_j)}^{\text{T}} \overrightarrow{\text{tf-idf}(d)} \tag{1}$$

If we consider the application of this method to the background corpus we can construct a matrix $\mathbf{X}$, whose elements are the corresponding TF-IDF values $x_{wj} = \text{tf-idf}_w(b_j)$, and hence that $\phi_i(b_j)$ is the i, j^{th} element of $\mathbf{X}^{\text{T}}\mathbf{X}$. One of the key assumptions is that we should have that topics that are as distinct as possible in order to reduce the amount of overlap between the topics. This is achieved by assuming that we have some function $sim : B \to [0, 1]$ that has the following property:

$$sim(b_i, b_j) = \begin{cases} 1 \text{ if } i = j \\ 0 \text{ if } i \neq j \end{cases} \tag{2}$$

This can be though of as maximizing training accuracy as we are ensuring that the similarity of two different topics in our background is zero and the similarity of the topic with itself is one. In McCrae et al. [4] it was shown that this can be achieved by the function[1]:

$$\Phi_{\text{ONETA}}(d) = (\mathbf{X}^{\text{T}}\mathbf{X})^{-1}\Phi(d) = \mathbf{X}^{+}\overrightarrow{\text{tf-idf}(d)} \tag{3}$$

[1] $\mathbf{X}^{+}$ denotes the Moore-Penrose pseudo-inverse, which satisfies $\mathbf{X}^{+}\mathbf{X} = \mathbf{I}$

For any choice of $\Phi(d)$ where $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$ exists and it is easy to verify that Equation 2 holds as:

$$\mathbf{I} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{X} = \mathbf{X}^{+}\mathbf{X} \tag{4}$$

In practice the computation of this matrix can be time-consuming so instead McCrae et al. proposed rearranging the order of the vocabulary in the background collection to find a good approximation of the form:

$$\mathbf{X} \simeq \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{0} & \mathbf{C} \end{pmatrix} \tag{5}$$

And as it is easy to verify the following equation based on a matrix of this form:[2]

$$\mathbf{I} = \begin{pmatrix} \mathbf{A}^{+} & -\mathbf{A}^{+}\mathbf{B}\mathbf{C}^{+} \\ \mathbf{0} & \mathbf{C}^{+} \end{pmatrix} \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{0} & \mathbf{C} \end{pmatrix} \tag{6}$$

This leads to a strong approximation of ONETA as follows:

$$\Phi'_{\mathrm{ONETA}}(d) = \begin{pmatrix} \mathbf{A}^{+} & -\mathbf{A}^{+}\mathbf{B}\mathbf{C}^{+} \\ \mathbf{0} & \mathbf{C}^{+} \end{pmatrix} \overrightarrow{\text{tf-idf}(d)} \tag{7}$$

3 Applying ONETA to dictionary inference

Language 1	Language 2	Articles	Words (Language 1)	Words (Language 2)
English	Spanish	521,286	546,995,808	346,483,264
English	French	575,795	599,662,371	376,420,982
English	Portuguese	273,331	349,277,387	151,379,113
French	Portuguese	180,060	173,208,723	122,155,565

Table 1. The size of the Wikipedia corpora used for our language pairs in terms of articles and number of words

The key purpose of ONETA is to estimate the similarity between documents, and to apply it to the task of inferring the similarity of translation, we make the simple assumption that each term in a translation is a single document consisting of only the term in question. As such we simply apply the system by building two ONETA functions for our source language, s, and target language t and estimate the similarity as:

[2] McCrae et al. use the Jacobi preconditioner of $\mathbf{C}$ as a further approximation of $\mathbf{C}^{+}$

$$sim(w_s, w_t) = \cos(\Phi^s_{\text{ONETA}}(w_s), \Phi^t_{\text{ONETA}}(w_t)) \tag{8}$$

In order to construct pairs, we considered only the simple pivot between one language, and as for this task the languages were English, French and Portuguese, there were only two common languages between them namely Spanish and Catalan, and as such we called our two systems ONETA-ES and ONETA-CA based on the pivot language. We simply considered all possible translations between the two language pairs and then calculated the similarity using the ONETA score. As we found that the distribution of the scores was strongly clustered around zero, we used the following function to provide a more even spread of scores.

$$sim'(w_s, w_t) = |sim(w_s, w_t)|^\alpha \tag{9}$$

For our experiments we tuned $\alpha = 0.3$ to provide a reasonable spread of certainty values. As with previous work we used Wikipedia to construct our corpora using the interlingual index to create a comparable corpus for each language pair, the sizes of which are given in Table 1.

4 Results

Threshold	Precision	Recall	F1
0.0	0.845	0.541	0.659
0.1	0.904	0.237	0.375
0.2	0.902	0.184	0.306
0.3	0.894	0.149	0.255
0.4	0.885	0.119	0.209
0.5	0.884	0.093	0.168
0.6	0.878	0.072	0.133
0.7	0.869	0.053	0.101
0.8	0.866	0.038	0.072
0.9	0.867	0.022	0.043

Table 2. Our results for translating from English to Spanish using Catalan as a pivot language

During development we evaluated on the English to Spanish translations using Catalan as a pivot, as all language pairs are available as part of the training data and the results are presented in Table 2. It should be noted that at the threshold value of 0% the system is basically nothing more than pivot translation and this should be considered a baseline. For higher values of the threshold, ONETA does improve the precision, however the recall also decreases rapidly causing the F-Measure to fall overall.

Threshold	Prec.	Recall	F1	Coverage	Prec.	Recall	F1	Coverage
0.0	0.65	0.22	0.33	0.39	0.61	0.30	0.40	0.52
0.1	0.69	0.19	0.30	0.34	0.66	0.25	0.36	0.43
0.2	0.74	0.17	0.27	0.29	0.71	0.21	0.32	0.36
0.3	0.78	0.14	0.23	0.23	0.75	0.17	0.28	0.3
0.4	0.81	0.10	0.18	0.18	0.79	0.13	0.23	0.23
0.5	0.83	0.08	0.14	0.13	0.81	0.10	0.17	0.17
0.6	0.85	0.05	0.10	0.09	0.83	0.07	0.12	0.12
0.7	0.87	0.03	0.06	0.06	0.84	0.04	0.08	0.08
0.8	0.87	0.02	0.04	0.03	0.85	0.03	0.05	0.05
0.9	0.88	0.01	0.02	0.02	0.86	0.01	0.02	0.02
1.0	0.81	0.00	0.00	0.00	0.79	0.00	0.00	0.00

Table 3. TIAD results for the ONETA system at various threshold values. The left side shows the pivot through Catalan and the right through Spanish

In the official results (Table 3), we see a similar outcome where the highest F-Measure is achieved at the trivial threshold of 0% and we see strong gains in precision at the cost of recall. This shows that ONETA can quite effectively select translations that are very likely to be correct but misses many translations even among those that are generated by a pivot method.

When all systems are compared (Figure 1) at various threshold levels we see that the ONETA-ES system actually reports the strongest F1 Measure (averaged over all language pairs) of any system, however it should be noted that this is a threshold value that we would consider to be a baseline. Even still, we see that at the threshold of 0.1, ONETA still has the second and eighth best result, moreover we have achieved the strongest precision scores across all languages (except for results with a recall that was reported as zero).

5 Conclusion

We have presented the ONETA system and its application to translation, which was the only system to produce a value that beat the baseline, albeit when it is in a mode where it effectively a baseline itself. The system does show noticeable ability to tune between precision and recall and as such it would likely be effective for usage in areas where precision is more important than recall, for example in a semi-automated setting where showing annotators too many poor quality translations would waste time. There are two principle flaws with the implementation as it stands: firstly, that the recall is limited and even our baseline mode we only achieved a recall of about 20-30%, which needs to be overcome by finding more translations than are just present in the graph. Secondly, the system is not aware of senses, and the selection of multiple document collections likely to show many different senses of a word, may help the system to distin-

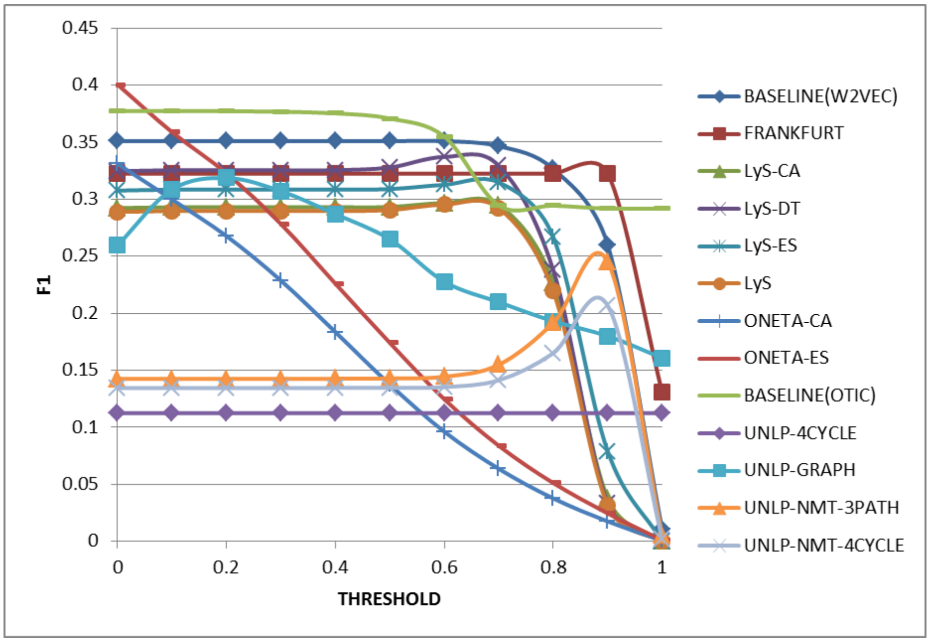

Fig. 1. F1 Results for all systems at various threshold levels, from the results of the TIAD task.

guish between translation pairs which do not rely on the most frequent senses of words.

Acknowledgment

This publication has emanated from research supported in part by a research grant from Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289, co-funded by the European Regional Development Fund, and the European Unions Horizon 2020 research and innovation programme under grant agreement No 731015, ELEXIS - European Lexical Infrastructure.

References

1. Aggarwal, N., Asooja, K., Bordea, G., Buitelaar, P.: Non-orthogonal explicit semantic analysis. In: Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics. pp. 92–100 (2015)
2. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet Allocation. Journal of machine Learning research **3**(Jan), 993–1022 (2003)
3. Gabrilovich, E., Markovitch, S., et al.: Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In: IJCAI. vol. 7, pp. 1606–1611 (2007)

4. McCrae, J., Cimiano, P., Klinger, R.: Orthonormal explicit topic analysis for cross-lingual document matching. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. pp. 1732–1742 (2013), https://www.aclweb.org/anthology/D/D13/D13-1179.pdf
5. Ordan, N., Gracia, J., Alper, M., Kernerman, I.: Proceedings of TIAD-2017 Shared Task – Translation Inference Across Dictionaries (2017)
6. Sorg, P., Cimiano, P.: An experimental comparison of explicit semantic analysis implementations for cross-language retrieval. In: International Conference on Application of Natural Language to Information Systems. pp. 36–48. Springer (2009)
7. Sorg, P., Cimiano, P.: Exploiting Wikipedia for cross-lingual and multilingual information retrieval. Data & Knowledge Engineering **74**, 26–45 (2012)
8. Vulić, I., Moens, M.F.: Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings. In: Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval. pp. 363–372. ACM (2015)