Unsupervised Cross-Lingual Scaling of Political Texts

Goran Glavaš and **Federico Nanni** and **Simone Paolo Ponzetto**
Data and Web Science Group
University of Mannheim
B6, 26, DE-68159 Mannheim, Germany
{goran, federico, simone}@informatik.uni-mannheim.de

Abstract

Political text scaling aims to linearly order parties and politicians across political dimensions (e.g., left-to-right ideology) based on textual content (e.g., politician speeches or party manifestos). Existing models scale texts based on relative word usage and cannot be used for cross-lingual analyses. Additionally, there is little quantitative evidence that the output of these models correlates with common political dimensions like left-to-right orientation. We propose a text scaling approach that leverages semantic representations of text and is suitable for cross-lingual political text scaling. We also propose a simple and straightforward setting for quantitative evaluation of political text scaling. Experimental results show that the semantically-informed scaling models better predict the party positions than the existing word-based models in two different political dimensions. Furthermore, the proposed models exhibit no drop in performance in the cross-lingual compared to monolingual setting.

1 Introduction

The goal of political scaling is to order political entities, i.e., political parties and politicians according to their positions in some political dimension (e.g., left vs. right ideological orientation). Textual content produced by political entities, such as parties' election manifestos or transcripts of speeches, is commonly used as the data underpinning the analyses (Grimmer and Stewart, 2013).

Advances in text mining have enabled various topical and ideological analyses of political texts. Computational methods for political text analysis cover dictionary-based models (Kellstedt, 2000; Young and Soroka, 2012), supervised classification models (Purpura and Hillard, 2006; Stewart and Zhukov, 2009; Verberne et al., 2014; Karan et al., 2016), and unsupervised scaling models (Slapin and Proksch, 2008; Proksch and Slapin, 2010). All of these models use the discrete, word-based representations of text. Recently, however, continuous semantic text representations (Mikolov et al., 2013b; Le and Mikolov, 2014; Kiros et al., 2015; Mrkšić et al., 2016) outperformed word-based text representations on a battery of mainstream natural language processing tasks (Kim, 2014; Bordes et al., 2014; Tang et al., 2016).

Although the idea of automated estimation of ideological beliefs is old (Abelson and Carroll, 1965), models estimating these beliefs from texts have only appeared in the last fifteen years (Laver and Garry, 2000; Laver et al., 2003; Slapin and Proksch, 2008; Proksch and Slapin, 2010). In the pioneering work on political text scaling, Laver and Garry (2000) used predefined dictionaries of words labeled with position scores. They then scored documents by aggregating the scores of dictionary words they contain. Extending this work, they proposed the model (Laver et al., 2003) that relies on manually labeled reference texts instead of dictionaries of position words. They then computed the lexical overlap between the unlabeled texts and the reference position texts.

Seeking to avoid the manual annotation effort, Slapin and Proksch (2008) proposed Wordfish, an unsupervised scaling model which has become the *de facto* standard method for political text scaling. Wordfish models document positions and contributions of individual words to those positions as latent variables of the Poisson naïve Bayes generative model, i.e., they assume that words are drawn independently from a Poisson distribution. They estimate the positions by maximizing the log-likelihood objective in which word variables inter-

Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, pages 688–693,
Valencia, Spain, April 3-7, 2017. ©2017 Association for Computational Linguistics

act with document variables.

In this work we aim to remedy for two major shortcomings pertaining to existing research on political text scaling:

(1) Existing methods rely on bag-of-words representations of text and are based on relative frequencies of words in documents being scaled. As such, they fail to exploit semantic similarities between words (e.g., *"bad hombre"* and *"terrible dude"* might indicate the same ideological position) and, more importantly, cannot be applied to cross-lingual scaling (i.e., scaling of texts written in different languages);

(2) Most existing studies provide only qualitative evaluation of the scaling quality and the extent to which automatically produced position scores correspond to actual positions of political actors.[1] Lack of transparent quantitative evaluation blurs insights into models' abilities to predict actual positions for a political dimension of interest.

The contributions of this paper are twofold. First, we propose an unsupervised scaling model which is, by exploiting semantic representations of text, equally suitable for monolingual and cross-lingual analyses of political texts. We exploit the recently ubiquitous word embeddings (Mikolov et al., 2013b; Pennington et al., 2014) to derive semantic representations of texts and the translation matrix model (Mikolov et al., 2013a) to construct a joint multilingual semantic vector space. We then build a fully-connected similarity graph by measuring semantic similarities between all pairs of texts. Finally we run a graph-based label propagation algorithm (Zhu and Goldberg, 2009) to derive final positions of political texts. Secondly, we propose a simple and straightforward quantitative evaluation that directly compares automatically produced positions with the ground truth positions (i.e., positions labeled by experts) for political dimensions of interest. Furthermore, we construct a dataset (with both monolingual and cross-lingual version), which we offer as a benchmark for quantitative evaluation of models for political text scaling.

2 Cross-Lingual Text Scaling

Our scaling approach consists of three components: (1) construction of a joint multilingual embedding space, (2) unsupervised measures of semantic similarity, and (3) a graph-based label propagation algorithm, which we use to derives final position scores from pairwise text similarities.

2.1 Multilingual Embedding Space

We start from monolingual word embeddings of all involved languages, obtained by running embedding models (Mikolov et al., 2013b; Pennington et al., 2014) on large corpora. Independently trained monolingual embedding spaces are in no way mutually associated, i.e., same concepts (e.g., English word *"bad"* and German *"schlecht"*) might have very different vectors.

In order to allow for semantic comparison of texts in different languages, we must construct a joint multilingual semantic vector space. To this end, we select the embedding space of one language and map embedding spaces of all other languages to the selected space using the linear translation matrix model of Mikolov et al. (2013a). Given a set of word translations pairs P, we learn a translation matrix $\mathbf{M}$ that projects embedding vectors from one embedding space to another. Let $\mathbf{S}$ and $\mathbf{T}$ be the matrices with monolingual embeddings of source and target words from P, respectively. Unlike the original work (Mikolov et al., 2013a), in which the matrix $\mathbf{M}$ is learned by numerically minimizing the differences between projections of source embeddings and target embeddings, we opt for a analytical solution for the matrix $\mathbf{M}$. Given that we want to find the matrix that translates $\mathbf{S}$ to $\mathbf{T}$, i.e., $\mathbf{S} \cdot \mathbf{M} = \mathbf{T}$ and that the source matrix S is not a square matrix (i.e., it does not have an inverse), we compute the translation matrix $\mathbf{M}$ by multiplying the pseudoinverse (inverse approximation for non-square matrices) of the source matrix $\mathbf{S}$ with the target matrix $\mathbf{T}$:

$$\mathbf{M} = \mathbf{S}^{+} \cdot \mathbf{T}$$

where $\mathbf{S}^{+}$ is the Moore-Penrose pseudoinverse of the source matrix $\mathbf{S}$, i.e., $\mathbf{S}^{+} = (\mathbf{S}^T\mathbf{S})^{-1}\mathbf{S}^T$. The translation matrices we obtained this way in our experiments turned to be of the same quality as those obtained via numeric optimization. However, the direct analytical computation using the pseudoinverse of the source matrix has the benefit of being significantly computationally faster than the numeric optimization.

[1] Proksch and Slapin (2010) perform a convolutedly indirect quantitative evaluation of Wordfish, which we do not find to be significantly more informative than qualitative evaluations.

2.2 Measures of Semantic Similarity

We propose two rather simple unsupervised measures of semantic similarity between texts that leverage the embeddings from the shared multilingual embedding space. Both similarity measures are fully language-agnostic, i.e., they simply use the joint embedding space to look up semantic vectors of words found in input texts.

Alignment similarity. The computation of the alignment score is based on the bijective alignment of words between two input texts. We *greedily* pair words between the two documents that have the most similar embedding vectors (according to the cosine distance) – once each word (more precisely, each token) has been aligned, it is not considered for further alignments. A similar alignment method has been proposed for evaluating machine translation systems (Lavie and Denkowski, 2009). Let t_1 and t_2 be the input texts and let $A = \{(w_1^i, w_2^i)\}_{i=1}^N$ be the obtained word alignment between them. The alignment similarity is then computed as follows:

$$s(t_1, t_2) = \frac{1}{N} \sum_{(w_1^i, w_2^i) \in A} \cos(e(w_1^i), e(w_2^i))$$

where $N = |A|$ is the number of aligned pairs, equal to the number of tokens in the shorter of the texts, and $e(w)$ is the embedding of the word w in the shared multilingual embedding space.

Aggregation similarity. Instead of aligning words of input texts according to their semantic similarity, aggregation score compares the aggregate semantic vectors of entire input texts. Let T be the bag of words of an input text t. We compute the aggregate embedding of the input text t as the sum of L2-normalized embeddings of words in T:

$$e(t) = \frac{1}{|T|} \sum_{w \in T} \frac{e(w)}{\|e(w)\|}$$

The aggregation similarity is then computed as the cosine of the angle between aggregate vectors of the two input texts:

$$s(t_1, t_2) = \cos(e(t_1), e(t_2))$$

2.3 Graph-Based Scaling Algorithm

With the shared embedding space and similarity metrics in place, we can compute semantic similarity scores for every pair of political texts we want to scale. The conversion of such pairwise text similarities into an one-dimensional scale of position scores is the final step of our scaling approach. Assuming that the two semantically most dissimilar texts, which we name *pivot texts*, represent the opposite position extremes for the political dimension of interest, we initially assign them extreme position scores of -1 and 1. Pairwise similarities between texts induce an undirected similarity graph and allow us to use graph-based score propagation to compute the positions for the remaining, *non-pivot* texts. Finally, after obtaining the positions of the non-pivot texts, we recompute the positions for the two pivot texts.

Position propagation. We use the *harmonic function label propagation* (HFLP)[2] (Zhu and Goldberg, 2009) – a commonly used graph-based algorithm for semi-supervised learning – to propagate position scores from the two pivot texts to other, non-pivot texts.[3] Before running the HFLP algorithm, we rescale all pairwise text similarities (i.e., all graph weights) to the $[0, 1]$ interval (i.e., 0 is the similarity between two least similar texts and 1 is the similarity between two most similar texts). Let $G = (V, E)$ be the similarity graph and $\mathbf{W}$ its weighted adjacency matrix. Let $\mathbf{D}$ be the diagonal matrix with weighted degrees of graph's vertices as diagonal elements, i.e., $D_{ii} = \sum_{j \in |V|} w_{ij}$, where w_{ij} is the weight of the edge between vertices i and j. The unnormalized Laplacian of the graph G is then given as $\mathbf{L} = \mathbf{D} - \mathbf{W}$. Assuming that the labeled vertices (in our case, the two vertices representing pivot texts) are ordered before the unlabeled ones, the Laplacian $\mathbf{L}$ can be partitioned as follows:

$$\mathbf{L} = \begin{pmatrix} \mathbf{L_{ll}} & \mathbf{L_{lu}} \\ \mathbf{L_{ul}} & \mathbf{L_{uu}} \end{pmatrix}$$

The harmonic function values of the unlabeled vertices, denoting the position scores of the non-pivot texts, are then given by:

$$\mathbf{f_u} = -\mathbf{L_{uu}^{-1}} \mathbf{L_{ul}} \mathbf{y_l}$$

where $\mathbf{y_l}$ is the vector of scores of labeled vertices, in our case, $\mathbf{y_l} = [-1, 1]^T$.

Rescaling pivot texts. We acknowledge that our two pivot texts (i.e., the pair of mutually least similar texts according to our semantic similarity measure) might not be the two texts expressing truly

[2] Also known as the *absorbing random walk*.

[3] Preliminarily, we also experimented with the PageRank algorithm (Page et al., 1999), but HFLP performed better.

the most dissimilar political positions because: (1) our metrics of semantic similarity are imperfect, i.e., the scores they produce are not the gold standard semantic similarities, but even if they were (2) we do not know to what extent the semantic similarity we measure correlates with the particular political dimension being analyzed (e.g., with the ideological left-to-right agreement). This is why, as the final step, we rescale the positions of the two pivot texts which we kept fixed for HFLP.

Let t be a pivot text and NP be the set of non-pivot texts for which we obtained the positions with HFLP. The final pivot text position is computed as the weighted sum of non-pivot positions:

$$p(t) = \sum_{t_i \in NP} p(t_i) \cdot s(t, t_i)$$

where $s(t, t_i)$ is the semantic similarity between texts t and t_i and $p(t_i)$ is the position of a non-pivot text t_i, obtained with HFLP. We finally rescale all position scores to range $[-1, 1]$, keeping the same proportions between pairs of party positions.

3 Evaluation

We first describe the dataset used for evaluation and then describe in detail the straightforward setting for quantitative evaluation of scaling methods. Finally, we interpret the obtained results.

3.1 Dataset

We collected a corpus of speeches from the fifth mandate of the European Parliament (EP) from the Parliament's official website. The choice of EP speeches for evaluation was a pragmatic one – each speech is available in all official EU languages, which allowed for a parallel monolingual and cross-lingual evaluation on the same set of speeches. We selected all speeches given by representatives from five largest European countries: Germany, France, United Kingdom, Italy, and Spain. We created aggregated texts for political parties by concatenating speeches of all party members. Finally, we kept the only the parties with aggregate texts longer than 15.000 tokens, which left us with a set of 25 political parties. We compiled the final dataset in the monolingual (English) and multilingual (speeches in speakers' respective native languages) versions.[4]

As in the previous work (Proksch and Slapin, 2010), we are considering party positions in two dimensions: (1) left-to-right ideology and (2) European integration. We obtained the gold party positions for both of these dimensions from the 2002 Chapel Hill expert survey.[5]

Source	Target	P@1 (%)	P@5 (%)
German	English	32.7	48.7
Spanish	English	46.6	58.3
Italian	English	34.4	52.5
French	English	36.4	56.2

Table 1: Evaluation of translation matrices.

3.2 Experimental Setting

Joint embedding space. We first obtain the monolingual word embeddings for all five languages in evaluation. We used the pretrained 200-dimensional GloVe word embeddings (Pennington et al., 2014) for English[6] and trained the 300-dimensional Word2Vec CBOW embeddings (Mikolov et al., 2013b) for the other four languages on respective Wikipedia instances. We induced the multilingual embedding space by translating embeddings of other four languages to the English embedding space. We obtained word translation pairs by translating 4200 most frequent English words to all other languages with Google translate. We used 4000 of the translation pairs to learn the translation matrices and remaining 200 for evaluation of translation quality. Translation quality we obtain, shown in Table 1 in terms of precisions at ranks one and five (P@1 and P@5), is comparable to that reported in (Mikolov et al., 2013a).

Models and evaluation metrics. We evaluate two different variants of our method, one employing the alignment similarity (ALIGN-HFLP) and the other computing the aggregation similarity (AGG-HFLP) for pairs of texts. We evaluate both models in both monolingual and cross-lingual scaling setting. For comparison, in the monolingual setting we also evaluate Wordfish (Slapin and Proksch, 2008). As a sanity check, we also evaluate a baseline that randomly assigns positions to texts.

Evaluation metrics. We use intuitive evaluation metrics for comparing model-produced positions with the gold positions: the pairwise accuracy (PA), i.e., the percentage of pairs with parties in the same

[4] We make the dataset and the scaling code available at https://bitbucket.org/gg42554/cl-scaling

[5] http://chesdata.eu/

[6] http://nlp.stanford.edu/data/glove.6B.zip

	Monolingual			Cross-lingual		
	PA	r_P	r_S	PA	r_P	r_S
Random	49.7	-.03	.00	49.7	-.03	.00
Wordfish	55.0	.21	.20	–	–	–
AL-HFLP	61.3	.35	.31	57.3	.20	.25
AGG-HFLP	**67.0**	**.53**	**.46**	**63.3**	**.34**	**.39**

Table 2: Scaling performance for the left-to-right ideological positioning.

	Monolingual			Cross-lingual		
	PA	r_P	r_S	PA	r_P	r_S
Random	49.1	.00	.00	49.1	.00	.00
Wordfish	59.7	.18	.33	–	–	–
AL-HFLP	**62.3**	**.25**	**.39**	**64.3**	**.54**	**.40**
AGG-HFLP	60.3	.24	.30	59.3	.48	.31

Table 3: Scaling performance for the positioning regarding European integration.

order as in the gold standard; and Spearman (r_S) and Pearson correlation (r_P) between the two sets of positions. While PA and Spearman correlation estimate the correctness of the ranking, Pearson correlation also captures the extent to which automated scaling reflects the gold distances between party positions.

3.3 Results and Discussion

In Tables 2 and 3 we show the models' scaling performance for two political dimensions – left-to-right ideology and European integration, respectively. Our semantically-aware models outperform the commonly used Wordfish model. For both dimensions, our best performing model significantly outperforms Wordfish ($p < 0.05$).[7] Positions produced by Wordfish seem to be better aligned with positions on European integration than with ideological left-to-right positions, which is in line with observations from (Proksch and Slapin, 2010). The same holds for our alignment model (ALIGN-HFLP). In contrast, the scaling based on the aggregation similarity measure (AGG-HFLP) seems to better correspond to the left-to-right ideological positioning. We hypothesize that this is because the comparison between semantically more imprecise aggregated text embeddings assigns more weight to the most salient dimension of speeches, which we speculate is the ideological position. In contrast, by comparing semantically more precise word embeddings, the alignment model treats all political dimensions of speeches more uniformly.

In the cross-lingual setting (i.e., when estimating positions from texts in different languages) we observe no (significant) drop in performance of our best performing model for either of the political dimensions with respect to the monolingual (English) setting. This crucial finding implies that our semantically-motivated approach for political text scaling is indeed as applicable to multilingual political corpora as it is to monolingual.

The performance levels that our models reach indicate that the semantic similarity scores we compute capture also similarities originating from dimensions other than the political dimension of analysis. For example, part of the similarity between parties from the same country comes from the mentions of the same country-specific issues (not mentioned by the parties from other countries), regardless of the ideological dis(agreement) between these parties. Because of these effects, we believe that text scaling models must be coupled with models that would previously extract only the portions of texts relevant for the dimension of analysis (e.g., a model for discerning ideological from non-ideological portions of text).

4 Conclusion

In this work, we presented what is, to the best of our knowledge, the first approach for cross-lingual scaling of political texts. We induce a multilingual embedding space and compute semantic similarities for all pairs of texts using unsupervised measures for semantic textual similarity. We then use a graph-based score propagation algorithm to transform pairwise similarities into position scores.

Experimental results from the straightforward quantitative evaluation we propose show that our semantically-informed scaling predicts party positions for two relevant political dimensions better than the commonly used Wordfish model. Moreover, the cross-lingual scaling performance of our models matches their monolingual performance, proving them to be suitable to scale political texts from multilingual collections.

We will next focus on cross-lingual classification models to pre-filter only relevant portions of text. Coupling such models with the presented scaling method will allow for measuring similarities only along the relevant political dimension (e.g., ideology) and lead to more accurate position estimates.

[7] According to the non-parametric stratified shuffling test (Yeh, 2000)

References

Robert P. Abelson and J Douglas Carroll. 1965. Computer simulation of individual belief systems. *The American Behavioral Scientist (pre-1986)*, 8(9):1–24.

Antoine Bordes, Jason Weston, and Nicolas Usunier. 2014. Open question answering with weakly supervised embedding models. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 165–180. Springer.

Justin Grimmer and Brandon M. Stewart. 2013. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, 21(3):267–297.

Mladen Karan, Daniela Širinić, Jan Šnajder, and Goran Glavaš. 2016. Analysis of policy agendas: Lessons learned from automatic topic classification of croatian political texts. In *Proceedings of the Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH) at ACL 2016*, pages 12–21.

Paul M Kellstedt. 2000. Media framing and the dynamics of racial policy preferences. *American Journal of Political Science*, 44 (2):245–260.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of EMNLP*, pages 1746–1751, Doha, Qatar, October. Association for Computational Linguistics.

Ryan Kiros, Yukun Zhu, Ruslan R. Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In *Proceedings of NIPS*, pages 3294–3302.

Michael Laver and John Garry. 2000. Estimating policy positions from political texts. *American Journal of Political Science*, 44 (3):619–634.

Michael Laver, Kenneth Benoit, and John Garry. 2003. Extracting policy positions from political texts using words as data. *American Political Science Review*, 97 (2):311–331.

Alon Lavie and Michael J. Denkowski. 2009. The meteor metric for automatic evaluation of machine translation. *Machine translation*, 23(2-3):105–115.

Quoc V. Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the International Conference on Machine Learning*, pages 1188–1196.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013a. Exploiting similarities among languages for machine translation. *CoRR*, abs/1309.4168.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Proceedings of NIPS*, pages 3111–3119.

Nikola Mrkšić, Diarmuid Ó Séaghdha, Blaise Thomson, Milica Gašić, Lina M. Rojas-Barahona, Pei-Hao Su, David Vandyke, Tsung-Hsien Wen, and Steve Young. 2016. Counter-fitting word vectors to linguistic constraints. In *Proceedings of NAACL*, pages 142–148. Association for Computational Linguistics.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. The PageRank citation ranking: Bringing order to the web. *Technical Report*.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Sven-Oliver Proksch and Jonathan B Slapin. 2010. Position taking in european parliament speeches. *British Journal of Political Science*, 40(3):587–611.

Stephen Purpura and Dustin Hillard. 2006. Automated classification of congressional legislation. In *Proceedings of the 2006 International Conference on Digital Government Research*, pages 219–225. Digital Government Society of North America.

Jonathan B. Slapin and Sven-Oliver Proksch. 2008. A scaling model for estimating time-series party positions from texts. *American Journal of Political Science*, 52(3):705–722.

Brandon M. Stewart and Yuri M Zhukov. 2009. Use of force and civil–military relations in russia: an automated content analysis. *Small Wars & Insurgencies*, 20(2):319–343.

Duyu Tang, Furu Wei, Bing Qin, Nan Yang, Ting Liu, and Ming Zhou. 2016. Sentiment embeddings with applications to sentiment analysis. *IEEE Transactions on Knowledge and Data Engineering*, 28(2):496–509.

Suzan Verberne, Eva Dhondt, Antal van den Bosch, and Maarten Marx. 2014. Automatic thematic classification of election manifestos. *Information Processing & Management*, 50(4):554–567.

Alexander Yeh. 2000. More accurate tests for the statistical significance of result differences. In *Proceedings of the Conference on Computational Linguistics (COLING)*, pages 947–953. Association for Computational Linguistics.

Lori Young and Stuart Soroka. 2012. Affective news: The automated coding of sentiment in political texts. *Political Communication*, 29(2):205–231.

Xiaojin Zhu and Andrew B. Goldberg. 2009. Introduction to semi-supervised learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 3(1):1–130.