Lost in Translation?

Evaluating the usefulness of machine translation for bag-of-words text models.

Erik de Vries
Martijn Schoonvelde
Gijs Schumacher

Abstract

Automated text analysis allows researchers to analyze large quantities of text. Yet researchers interested in comparative questions are presented with a big challenge: across countries people speak different languages. To address this issue, some analysts have suggested to use Google Translate to convert all texts to English before starting the analysis (Lucas et al., 2015). But in doing so, do we get lost in translation? This paper evaluates the usefulness of machine translation for automated bag-of-words models – such as topic models – using the europarl dataset (which contains the transcriptions of professional translators of debates in the European Parliament in English and in the original language) to compare the output of a topic model of hand-translated data with a topic model of machine-translated data. To this end, we analyze at the similarities at the document-level and and the topic-level. We find that there are only small differences between the topic models of hand-translated texts and machine-translated texts. Specifically, at the document level we find that over 92% of document pairs achieve a cosine similarity of 0.80 or higher. At the topic level, we find topic distributions to be highly correlated: 65% of document pairs have a correlation of 0.80 or higher. Furthermore, we find only marginal differences between languages. We conclude that Google Translate can be a useful tool for researchers who use text to analyze comparative questions.
Keywords: automated text analysis, google translate, bag-of-words

Introduction

Automated text analysis is like a gold rush. Lots of researchers have noticed its potential and are now using methods such as topic modeling, scaling and sentiment analysis to address questions about consumers, political elites and individuals' mood (for an overview see Grimmer and Stewart, 2013). But much like the gold rush didn't make everyone rich, users of automated text analysis methods are starting to feel their shortcomings as well. In particular, when comparing text across countries, analysts face the simple but harsh reality that people speak different languages and it cannot be assumed that automated text analysis methods can directly cope with this. Thus, in order to make comparisons across countries the analyst first needs to translate texts from several languages into one. On the plus side, this can be automated by using machine translation, like, for example, Google Translate. But do texts get lost in Google Translation? That is, do we lose (too much) information if we first Google Translate texts and then analyze them? Or does applying automated text analysis methods across countries leave us like the poor souls who journeyed west for gold but left with nothing?

This paper evaluates the usefulness of machine translation for automated bag-of-words models, comparing bag-of-words vectors of machine translated and gold standard texts. By doing so, the approach of this paper fundamentally differs from those most commonly used in computer science, as computer science research concerned with machine translation strongly focuses on the more abstract concept of quality estimation(i.e. Scarton and Specia (2014), Kaljahi and Samad (2015), Aharoni (2015)), which means as much as programatically evaluating the quality of machine-translated texts without the need for a manually translated gold standard. In contrast, this paper, rather than purely evaluating the linguistic output of machine translations, seeks to evaluate machine translations in a more practical (social science) research context.

To do so, we use the europarl dataset (Koehn, 2005) which contains the official transcriptions of debates in the European Parliament both in English and in their original language. From this dataset we take debate contributions in Danish, German, Spanish, French and Polish from the period January 2007 to November 2011. Coming from professional translators, these official transcriptions serve as our gold standard.[1] Then we move to an application of such a bag-of-words model, topic modeling, to analyze if and how topics get lost in Google translation. Specifically, at the level of documents we analyze the correlations between the two topic distributions, and at the level of topics we analyze the distribution of words.

We find only small differences between topic models of hand-translated texts and machine-translated texts. Specifically, across languages we find that over 92% of document pairs achieved a cosine similarity of 0.80 or higher, and although we find significant differences across languages, their effect sizes are small indeed. What is more, we find high overlap in the set of features between hand-translated and machine-translated texts. At the topic level, we find topic distributions to be highly correlated: 65% of document pairs have a correlation of 0.80 or higher. Furthermore, on average 58% (equal number. of topics) or 60% (unequal number of topics) of topic pairs correlate 0.80 or higher in their distribution over all documents.

This paper proceeds as follows. We first briefly review the bag-of-words approach, topic models and machine translation theory. We then present our data, analysis and results. We end with a conclusion.

Background

To get at the topical content of the speeches contained in the europarl data, we apply Latent Dirichlet Allocation (LDA) models to our sample of speeches. LDA is based on the assump-

[1] Costs of hiring professional translators in the European Union are high, by some estimates as high as €2 per EU inhabitant per year (see `http://ec.europa.eu/dgs/translation/faq/index_en.htm`).

tion that each text (in our case: speech) can be represented as a probability distribution over topics, where topics themselves are considered to be a probability distribution over words (Blei, 2012; Blei, Ng and Jordan, 2003; Boumans and Trilling, 2016). LDA is a generative model, which takes the words in each document as input, and from that input plus some statistical assumptions infers the hidden topical structure—the topics, per-document topic distributions, and the per-document per-word topic assignments. Since LDA is an unsupervised method, the meaning of a topic—probability distributions over words—has to be inferred by the researcher.[2]

While numerous studies analyze machine-translated texts, little is still known about the quality of these translations and how they may affect the results of subsequent analyses. Usually, authors either assume machine-translated text to be suitable for their purposes or they do not pay attention to the issue altogether. For example, Agarwal et al. (2011) use Twitter data supplied by an unidentified commercial source which was – at least partly – translated using Google Translate. However, besides mentioning this, they make no remarks on the possible influence of (partially) machine-translated data on their analyses and results. Similarly, Benoit, Schwarz and Traber (2012) use Google Translate in the multilingual Swiss context. While these authors do describe how their translations were carried out, and which translation strategy yielded the best results (in this case, translating all languages to English), they give no details on the comparisons between different translation strategies, nor on the impact of translation on the results. This goes to show that much can still be learned about the impact of machine translations on specific analyses and research designs.

However, the above does not imply that machine-translation is not a viable strategy for analyzing texts in multiple languages. As Lotz and Van Rensburg (2014) show, developments in machine-translation systems are going fast, and their quality is clearly increasing over

[2]This paper is concerned with evaluating machine-translation quality; substantive interpretations of the topics are beyond its scope.

time. Besides that, Balahur and Turchi (2014) give a more comprehensive account of the use of machine-translated text for automated analyses, but do so in the context of sentiment analysis instead of topic modeling. Also, while discussing many details and methods, they do not discuss the implications of machine-translation for bag-of-words methods (which sentiment analysis is). The same is true for Lucas et al. (2015), who describe extensively the possible pitfalls of using machine-translated text in automated text analyses. However, they do not evaluate the performance of machine-translated text in bag-of-words analyses. In contrast, this paper evaluates both the bag-of-words approach in general and LDA topic modeling in particular.

There is one other topic that is relevant to the current study, and that is the impact of specific languages and language groups on machine translation quality. For example, machine-translated texts may be of better quality when translating from French to English than when translating from Polish to English. To examine this, the choice has been made explicitly to include languages from different language groups. These are French and Spanish (belonging to the Italic language group), German and Danish (belonging to the Germanic language group), and Polish (belonging to the Balto-Slavic language group). While there are different topologies of language groups to be found in the field of linguistics, this paper uses the topology as described in Gray and Atkinson (2003). The differences between these language groups are of importance as some language pairs can be more easily translated to each other than others (Koehn and Monz, 2006). In addition, for some languages larger parallel corpora are available to train machine-translation models on than others (e.g. there is more parallel data available for French and English than there is for Polish and English). Thus, machine-translation models might be better equipped to translate from and to some language pairs than to others (Lucas et al., 2015). Based on both arguments, it is reasonable to assume that there might be differences in translation quality between the languages included in this study, and thus also in the usability of machine translation for

bag-of-words approaches. Therefore, we will pay particular attention to the question if there are statistically significant differences between the results of different languages.

Data & Measurement

To evaluate the usefulness of using machine-translated texts for automated bag-of-words text analyses, we need to compare them against a gold standard. We thus require parallel corpora, the documents on which the evaluation is done need to be available in both English and the original language. The europarl dataset (Koehn, 2005) contains parallel corpora. This dataset contains the official transcriptions of debates in the European Parliament in most of the official EU languages from April 1996 until November 2011.

However, some countries became member states at a more recent date than April 1996. Most notably, countries that have a Slavic language are only part of the dataset from January 2007 onwards. To keep the time frame identical for all languages, including Polish, only data from 2007 until 2011 is analyzed.

The primary purpose of the europarl dataset is to train, test and improve machine translation algorithms (e.g. Koehn, 2005; Popescu-Belis et al., 2012; Loaiciga, Meyer and Popescu-Belis, 2014), and, because of that, the primary form in which the data is available is through text files with sentence-aligned language pairs. In these files, however, no distinction is made between different sessions of parliament, and thus no distinction can be made between different documents on which a topic model might be run. Therefore, the raw data provided by the project has been processed manually. The raw data is provided in the form of sessions of parliament (covering a single day), which can then be subdivided into chapters concerning different debates, questions and votes. However, the provided data is not exactly the same for all languages (e.g. chapter 5 in the session of 04-01-2007 might be present in the English but not in the German data, while the German data does contain other chapters from that

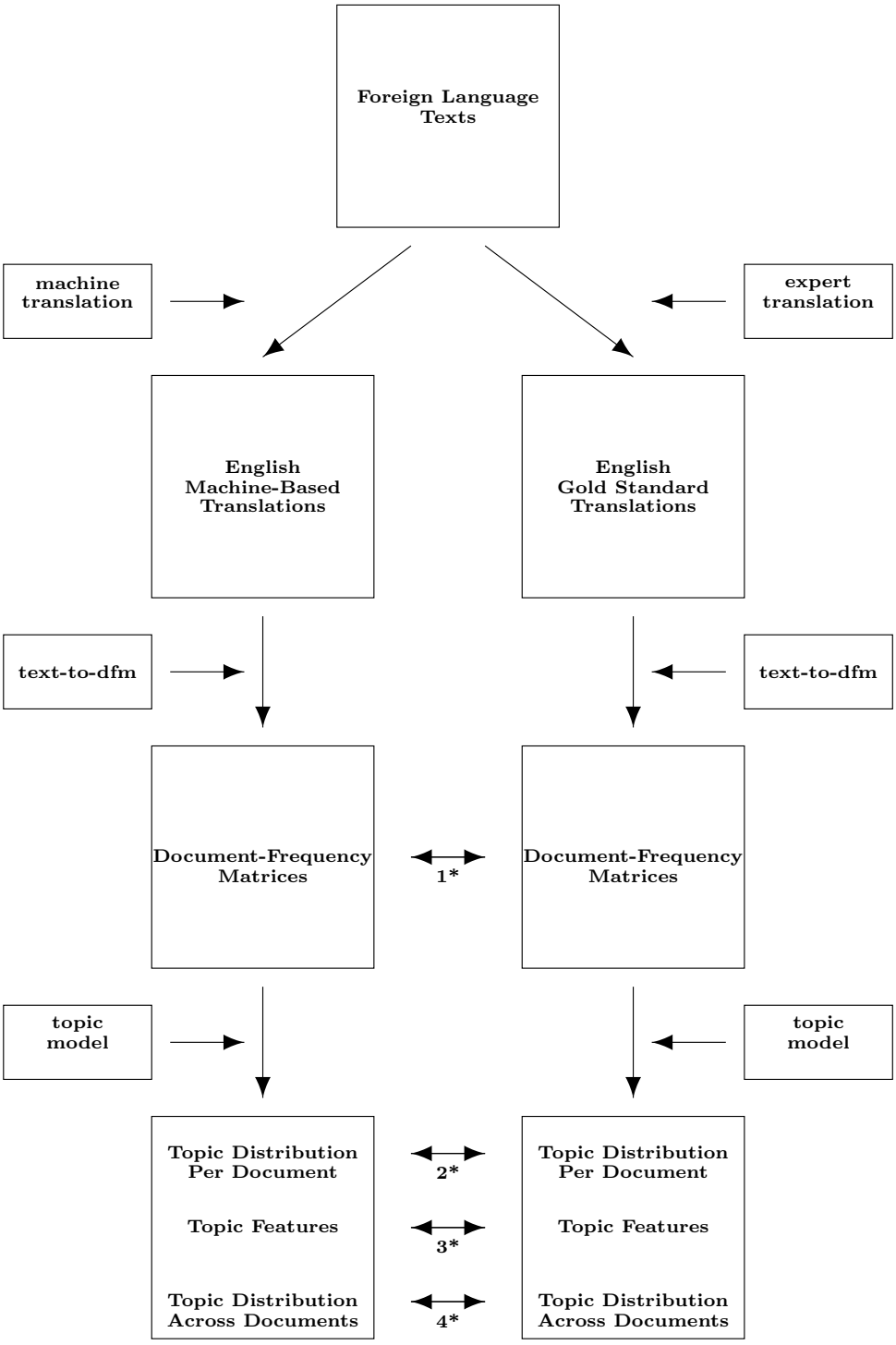
same session). Therefore, all language pairs (EN-DA, EN-DE, EN-ES, EN-FR, EN-PL) have been matched by checking for the presence of each chapter in each session for both languages. This results in between 2148 (DE) and 2347 (FR) chapters per language pair. In the topic models, each of these chapters is considered to be a single document, because the different chapters in a session concern substantively different subjects.

Methods

Google Translate

We use Google Translate as the specific machine-translation service to evaluate the performance of machine-translated texts in bag-of-words analyses. We chose Google Translate because of its translation quality, which compared to other online machine translating services is top-tier (Hampshire and Salvia, 2010).

Figure 1: What are we comparing?



Note:
1*) Document-to-document similarity: comparison of similarity of dfms.
2*) Document-to-document topic distribution: Are the distributions of topics over the document the same for both documents?
3*) Topic-to-topic similarity: Do the shared features (stems that occur in both dfms) have the same weight for the topic pair?
4*) Topic-to-topic distribution: How does the distribution of a single topic over all documents compare between models?

Figure 1 shows how we compare machine-based and expert translated documents. In both cases we start with identical non-English texts, which have been translated into English, either through Google Translate or through EU-employed expert translators. The English translations are then turned into bags-of-words (i.e., document-frequency matrices) on which we then estimate a topic model. To evaluate the usefulness of Google Translate we compare features of the models estimated from the Google Translate translations and the gold standard translations from EU-experts.

Pre-processing

When using bag-of-words models, it is common to pre-process the data in order to remove noise. These pre-processing steps might have a large impact on the outcome of automated text analyses (Denny and Spirling, 2016; Greene et al., 2016). In our case we only take some minimal pre-processing steps that are of little consequence. We have removed punctuation and general stopwords, and all remaining words have been lowercased and stemmed. Removal of punctuation should have no discernible impact on bag-of-words approaches, as word order within a document is irrelevant. Lowercasing could have had some impact, for example Rose (name) and rose (flower) might have been merged (Denny and Spirling, 2016). The removal of stopwords would not have had much impact, as these words occur frequently, and do not relate specifically to any topic.

In order to perform these pre-processing steps, we used both Python and R packages. For stemming, stopword removal, lowercasing and punctuation removal we used regular expressions in Python and the *NLTK* package (Bird, Klein and Loper, 2009). To create the actual bag of words (document-frequency matrices) we switched to R and the *quanteda* package (Benoit and Nulty, 2013). This package also removes numbers. Please note that the pre-processing steps on both the gold standard and machine translated texts are identical, and pre-processing of the machine translated texts was conducted after translation; we fed

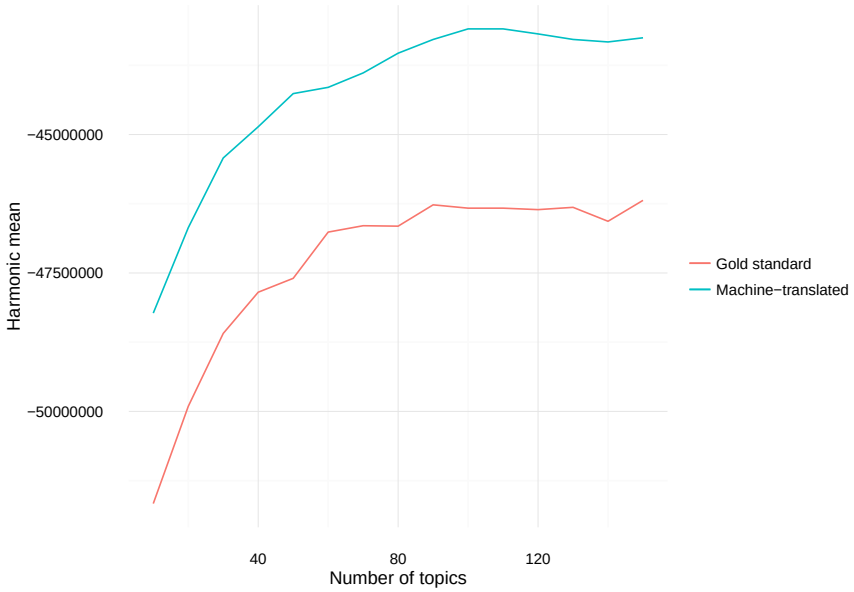
the machine translation algorithm the original texts with no edits.

Topic modeling

To assess the quality of machine translated texts in bag-of-words automated text analysis methods, we estimated topic models for the hand-translated and machine-translated texts using the LDA algorithm (Blei, Ng and Jordan, 2003). For this we used the LDA function in the *topicmodels* package in R (Hornik and Grün, 2011). In order to make sure that differences between the model based on the gold standard corpus and the model based on the machine-translated corpus are solely the result of differences between these corpora, all other relevant factors were kept constant. Most importantly, the number of topics was kept constant, and a fixed seed was used – based on the sys.time variable – as suggested by Hornik and Grün (2011). Other factors, such as the number of (burnin) iterations were also kept constant. Furthermore, we tested the stability of the models by running a topic model with identical parameters twice on the same data, which resulted in identical models. Consequently, all variation between the models – when the model parameters are kept the same – results from differences in the data itself.

Because it was not feasible to run and optimize the number of topics for each language pair, and because all language pairs are based on roughly similar data from the same time period, the optimum number of topics for all models was determined based on the French dataset, the largest gold standard and machine-translated dataset. Optimization of this corpus has been done by evaluating the model harmonic mean of models that contain between 10 and 150 topics, in increments of 10. Model harmonic mean indicates the extent to which word counts in the documents used to construct the model match the word distributions in the model itself. Or, put differently, to what extent the model accurately describes the distribution of words in the documents. The results of the optimization runs are displayed in figure 2.

Figure 2: Model harmonic mean



Choosing the right (optimum) number of topics for any given model is very important, as the number of topics determines the distribution of words over topics and, as a result, the distribution of topics over documents. When the number of topics changes, so do these distributions, and, consequently, the outcome of the model in its entirety. Therefore, in order to only detect the pure effect of machine translation on topic models, an argument can be made to keep the model parameters in both models exactly the same, including the number of topics. However, the plot above shows that the optimum number of topics (the point where the harmonic mean starts to decrease for the first time) is not the same for the gold standard and machine-translated models. While quite close, the gold standard model has an optimum of 90 topics, while the machine-translated model has its optimum at 100 topics. Combined with the arguments above, a choice can be made both to run all models with the same number (90) of topics, or to run the gold standard models at 90 topics, and the machine-translated models at 100 topics. But most importantly, in practical applications it is unlikely that machine translations will be used alongside gold standard translations.

Therefore it is likely that the optimum number of topics will often be chosen based on only the machine translations. On the other side, keeping the number of topics constant between models enables the evaluation of the effect of machine translation on the models. Therefore comparisons will be made both between gold standard and machine-translated models with an equal number of topics (90), and between models which have a number of topics that fits their own optimum (90 topics for gold standard models, 100 topics for machine-translated models). Comparing the results of those between-model comparisons also makes it possible to get an indication to what extent running models at their optimum number of topics influences the comparability of gold standard versus machine-translated models.

Matching topics

Now that we have chosen an appropriate number of topics for the models, our next challenge is to match the topics generated by the gold standard and machine-translated models to each other, because only in this way can we compare the results. This matching has been done by taking all dfm features (word stems) that are present in both the machine-translated and gold standard dfm's, and then determine for each of those features on which topic it loads highest. This results in a topic pair for each dfm feature. Summing the occurrences of each topic pair and sorting the resulting list in descending order of occurrence leads to a simple way of determining what topics belong together, and how strongly they do. To avoid assigning any one topic to more than one topic in the other model, topic pairs are generated from this list in order from the most to the least occurrences, but also by checking for each pair if one of the two topics is already assigned to another topic. If one of the two topics is already assigned, that specific topic pair is discarded. In this way, topic pairs are always the two topics that most strongly belong together, without assigning any one of those topics more than once to another topic.

However, there are also topics for which the highest loading features load higher on

another topic from the same model. As such, those topics could not be matched to any topic in the other model. In general however, this also means that the topics that could not be matched are of relatively little importance, as the most important features that compose them are by definition more important to other topics. In the same way, also superfluous topics in the comparison between models with different numbers of topics are dropped, and again the topics that are dropped are by definition the ones that are the most weakly linked to any of the topics in the other model.

For the comparison between models with 90 and 100 topics, 10 topics from the 100 topic model were dropped from the model in all languages, while only for Polish it was also necessary to drop an additional topic from both the 90- and 100-topic model, because they could not be matched. For the comparison between models with both 90 topics, for all languages except German 89 topics could be matched to each other, which means for those languages 1 topic in both the gold standard and machine-translated model was dropped. For German, all 90 topics could be matched to each other, and thus no topics were dropped in this case. The reason that some topics could not be matched to each other, is because the most important features (stems) identifying them either already loaded higher on another topic, or could not be used to link the two together. As those topics obviously by definition have very little in common, they were dropped from the data.

Comparing results

In order to evaluate the usability of machine-translated text for automated (bag-of-words) text analysis methods, we will compare throughout the results section between gold standard and machine-translated data. We will structure the results around four different comparisons, for which two aspects are most relevant. First, the difference between similarity and distribution, where the former is concerned with the extent to which dfm features (stems) actually match between gold standard and machine-translated data, and the latter is con-

cerned with the comparability of the topic model outcome (the distribution of topics over documents). Second, the distinction between comparisons on the document and the topic level, where the former is concerned with how individual document pairs compare to each other, while the latter is concerned with how topic pairs compare to each other. This twofold distinction results in the comparison matrix presented in Table 1.

Table 1: Comparisons between gold-standard and machine-translated data

Similarity	*Distribution*
Dfm features per document pair	Topic distribution per document pair
Word loadings per topic pair	Per-topic pair distribution over all documents

In addition to these comparisons, similarity measures also differ between different comparisons, most notably between the dfm comparisons and the topic model comparisons. For the dfm comparisons, we use cosine similarity as a measure because in contrast to correlation it takes into account the absolute differences in values between two documents, which is of importance when comparing dfm's because our goal of knowing how close the counts of all dfm features per document pair are to each other. For topic model comparisons, however, correlation is a more suitable similarity measure, because it compares trends rather than absolute values. The reason for this is that comparisons will be made between models with a different number of topics. As already discussed above, changing the number of topics influences both the topic-over-document and word-over-topic distributions, and because of that absolute values can no longer be compared.

Results

Comparing the bag of words

A first step in evaluating the usability of machine-translated text for bag-of-words models is to actually compare the two bags-of-words to each other. This can be done by comparing so-called document-frequency matrices (dfm's) of the gold standard and machine-translated documents using the built-in similarity function in the *quanteda* R package. Figure 3 shows the distribution of cosine similarity scores for each language. Most notably, the average similarity between the gold standard documents and their machine-translated counterparts is high (M=0.92, SD=0.07). This also shows in more than 92% of all document pairs achieving a cosine similarity score of .8 or higher.

Figure 3: Distribution of cosine similarity per language pair

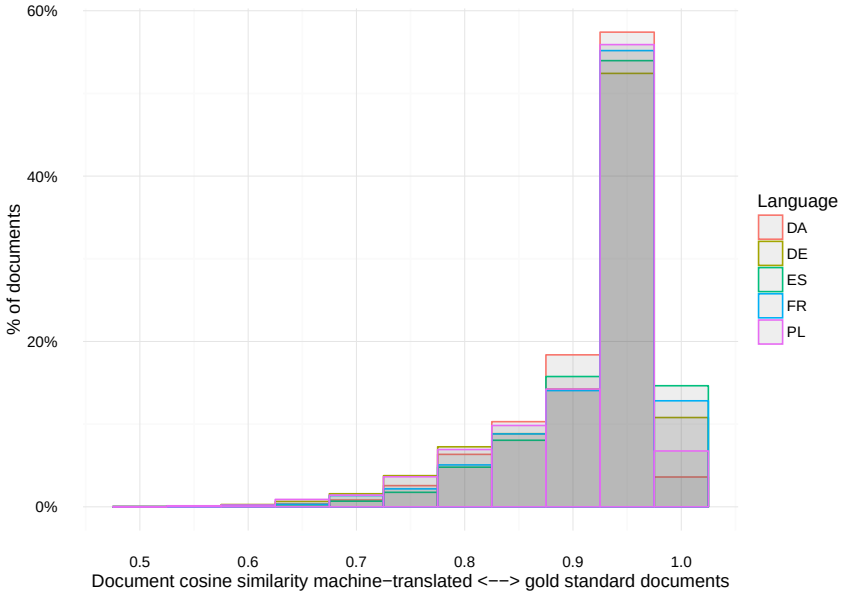


Table 2 shows means and standard deviations for document cosine similarity scores of each language. Note that while there are statistically significant differences between the languages, effect sizes are marginal. The significance of the differences is caused by the

French and Spanish data, which are the only two groups within the data that have significant t-values (French: t=7.07, p<0.001; Spanish: t=5.11, p<0.001)

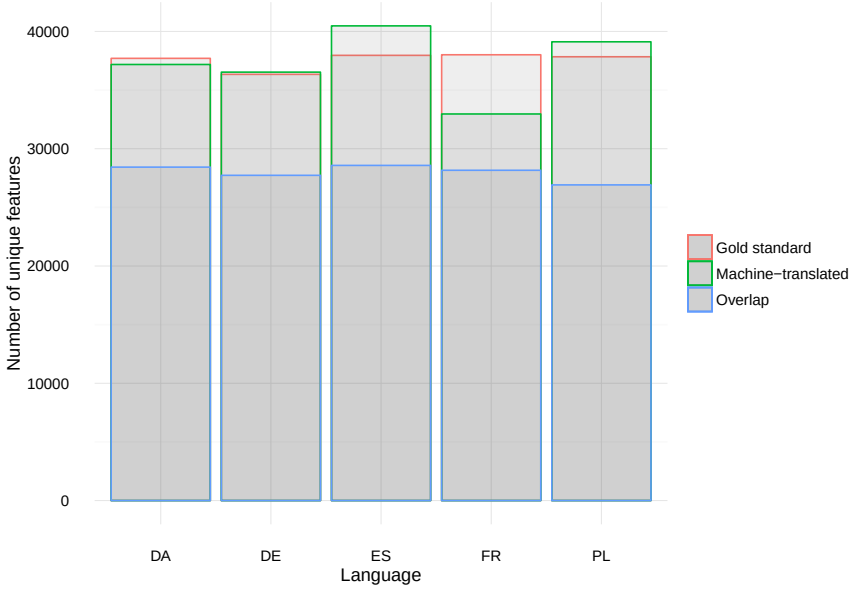
Table 2: Cosine similarity distribution per language

Language	N	Mean	St. Dev.	Min	Max
Danish	2,301	0.915	0.063	0.549	0.992
German	2,148	0.915	0.074	0.488	0.991
Spanish	2,335	0.929	0.059	0.483	0.991
French	2,347	0.925	0.064	0.564	0.989
Polish	2,338	0.913	0.073	0.475	0.989
Total:	11,469	0.920	0.067	0.475	0.992

Note: Statistically significant but very small difference between languages (ANOVA results: $F(4, 11464) = 27.855$, $\rho < 0.001$, $\eta^2 = 0.010$)

Another measure of how strongly gold standard and machine-translated dfm's differ from each other is the difference in the total number of unique features between the two, as well as the number of features that are shared between them. The overlap and differences between gold standard and machine-translated dfm's for each language are presented in figure 4.

Figure 4: Unique dfm features for gold standard and machine-translated corpora



Reading example: For French, the amount of overlapping features is around 28,000, while the total number of features is around 32,000 for the machine-translated documents and around 38,000 for the gold standard documents.

Overall, these statistics show that the document frequency matrices of the gold standard and machine-translated documents overlap to a substantial extent in terms of unique features. The number of overlapping unique features is also quite constant between the languages. The same goes for the unique features that are specific to either the gold standard or machine translated dfm's. This is with the notable exception of French, and to a lesser extent also Spanish, which are also the two languages responsible for the significant differences in the dfm cosine similarity comparison. While this study is not aimed at explaining differences in machine-translation performance in different languages, it seems that Google Translate has more trouble with translating these languages to English.[3] In the Spanish case, more unique features are present in the machine-translated than in the gold standard texts, which

[3]It would be very interesting to see if these unique features are actually caused by inaccurate, but in meaning similar translations. However, due to the automated nature of all the analyses conducted here, this is not within the scope of the current paper.

indicates that Google Translate disproportionally adds new features to the texts (by using different English translations for the same Spanish word). Similarly, French translations are disproportionally simplified (different French words translated as the same English word), because there are notably less unique features specific to the machine-translated documents than there are in the gold standard documents. Overall, however, there is substantial overlap between the unique features, and in combination with the very high cosine similarity scores of individual document pairs in the dfm's there are clear preliminary indicators that Google Translate is usable in bag-of-words approaches.

How do topic models compare?

Figures 5 and 6 show the distribution of correlation scores for all documents in the gold standard and machine-translated corpora. The correlation scores in this case indicate the extent to which the distribution of topics over each document matches between the two corpora. Or in other words, the extent to which the gold standard and machine-translated documents are characterized in the same way by their respective models.

Figure 5: Document topic distribution correlation with equal no. of topics

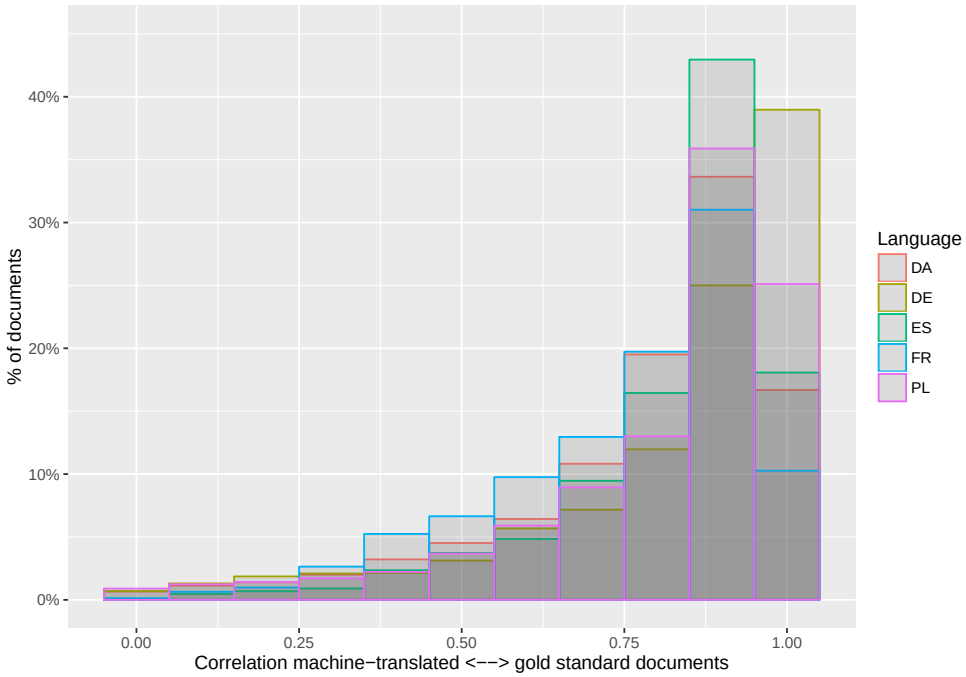


Figure 6: Document topic distribution correlation with unequal no. of topics

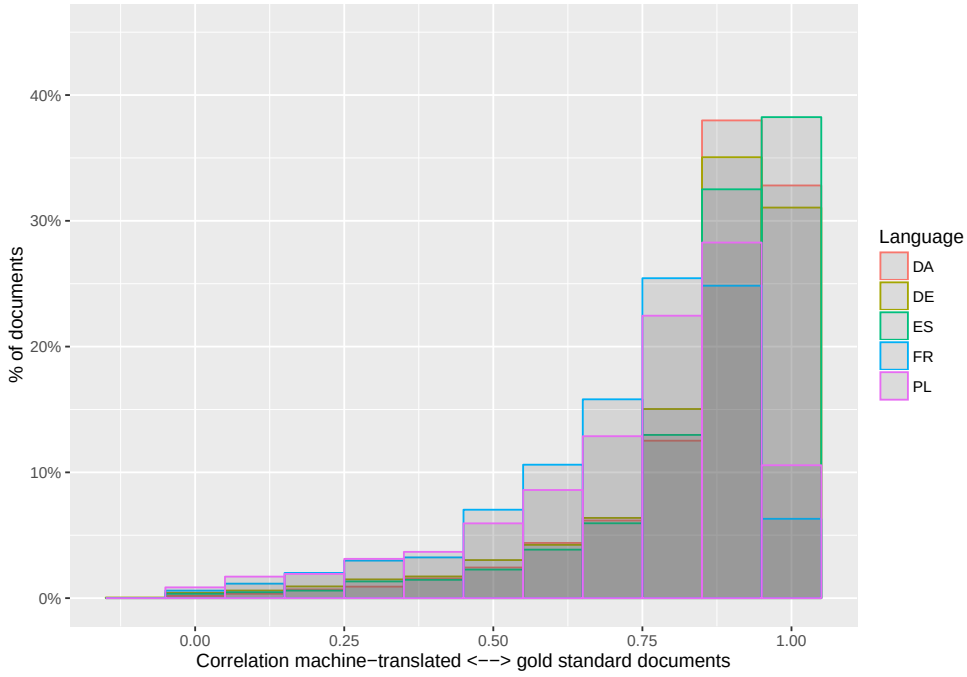


Figure 5 shows the resulting correlation between documents when the number of topics for the gold standard and machine translated documents is kept equal (90 topics per model), while figure 6 shows the same statistics, but with a different number of topics for the gold standard (90) and the machine-translated (100) models. Besides the obvious notion that there is a very small number of documents which actually have a negative correlation to each other in the comparsion of unequal models, overall the distribution of scores in the comparison of unequal models is quite comparable to the distribution of scores in the equal model comparison. This shows most importantly that changing the number of topics towards the optimum for both datasets does not severely influence document correlation scores. But it also shows that topic distributions per document are highly similar between gold standard and machine translated models, with on average – over all languages – 65% of document pairs having a correlation of 0.8 or higher. However, that being said, there are statistically significant differences between languages, even though their effect sizes remain rather small. Note that t-values for all languages are significant regardless of comparing equal or unequal numbers of topics, so the statistical significance of differences cannot be attributed to a specific language.

Table 3: Document topic distribution correlation with equal no. of topics

Statistic	N	Mean	St. Dev.	Min	Max
DA	2,301	0.783	0.202	−0.031	0.998
DE	2,148	0.824	0.216	−0.031	0.999
ES	2,335	0.826	0.165	0.028	0.997
FR	2,347	0.753	0.194	−0.043	0.996
PL	2,338	0.809	0.206	−0.031	0.998
Total	11469	0.799	0.199	−0.053	0.999

Note: ANOVA results: $F(4, 11464) = 56.414$, $\rho < 0.001$, $\eta^2 = 0.019$

Tables 3 and 4 show the mean correlation of topic distributions between gold standard and machine-translated documents from the same language, as well as the total number of

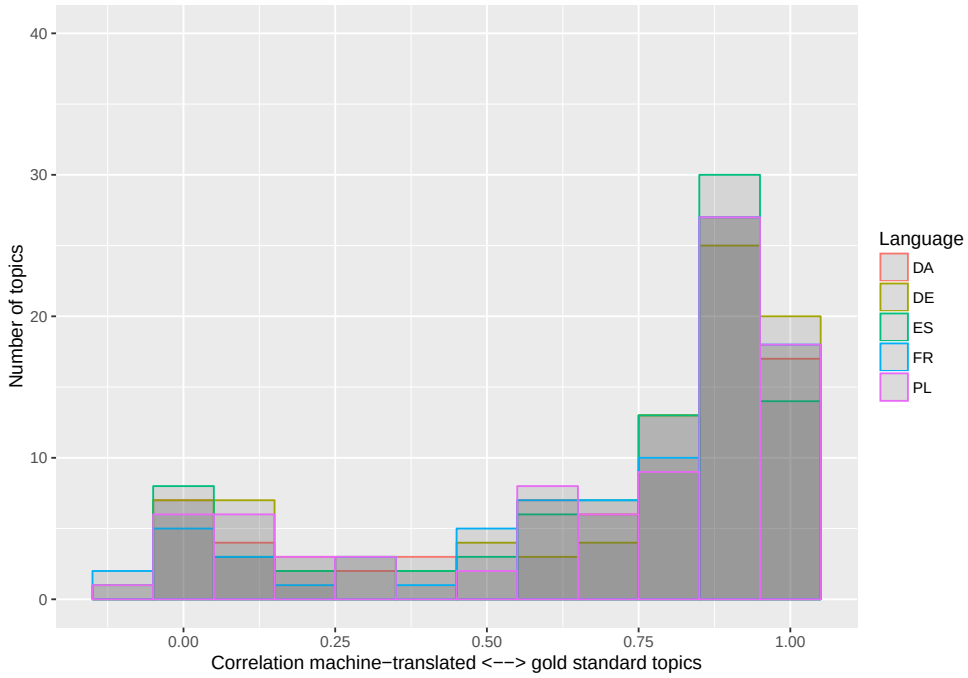
documents and standard deviations. Table 3 provides descriptives of comparisons between models with both 90 topics, while table 4 shows those for the comparisons of models with an unequal (90 vs. 100) number of topics.

Table 4: Document topic distribution correlation with unequal no. of topics

Statistic	N	Mean	St. Dev.	Min	Max
DA	2,301	0.859	0.161	−0.039	0.998
DE	2,148	0.842	0.181	−0.051	0.998
ES	2,335	0.860	0.168	−0.030	0.998
FR	2,347	0.727	0.201	−0.047	0.998
PL	2,338	0.740	0.216	−0.035	0.994
Total	11469	0.805	0.196	−0.051	0.998

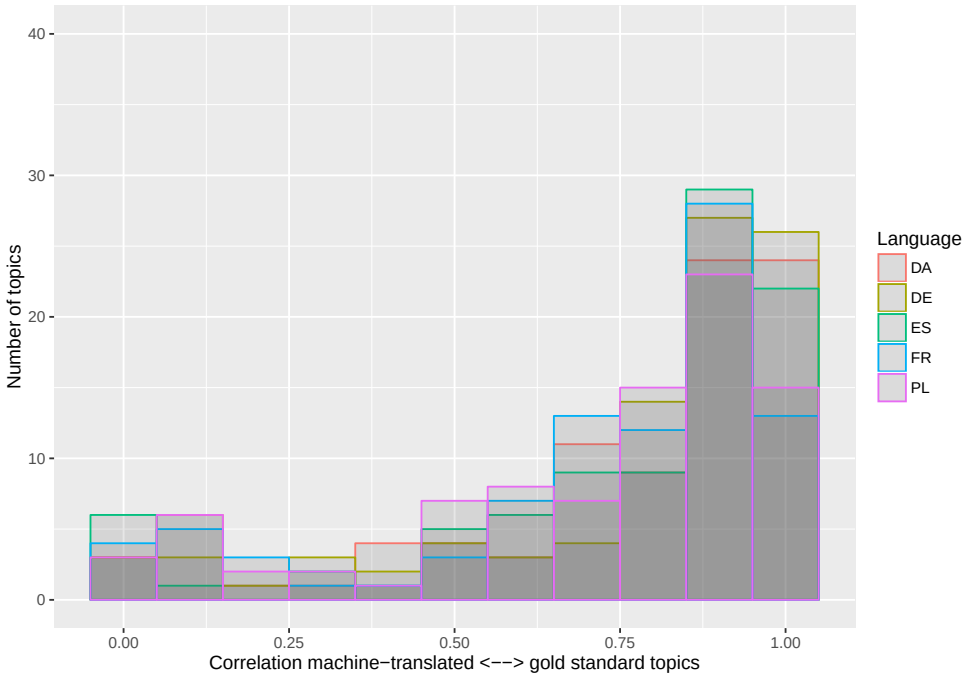
Note: ANOVA results: $F(4, 11464) = 294$, $\rho < 0.001$, $\eta^2 = 0.093$

Figure 7: Single topic distribution correlation with equal no. of topics



Overall descriptives: N=446, M=0.699, SD=0.321

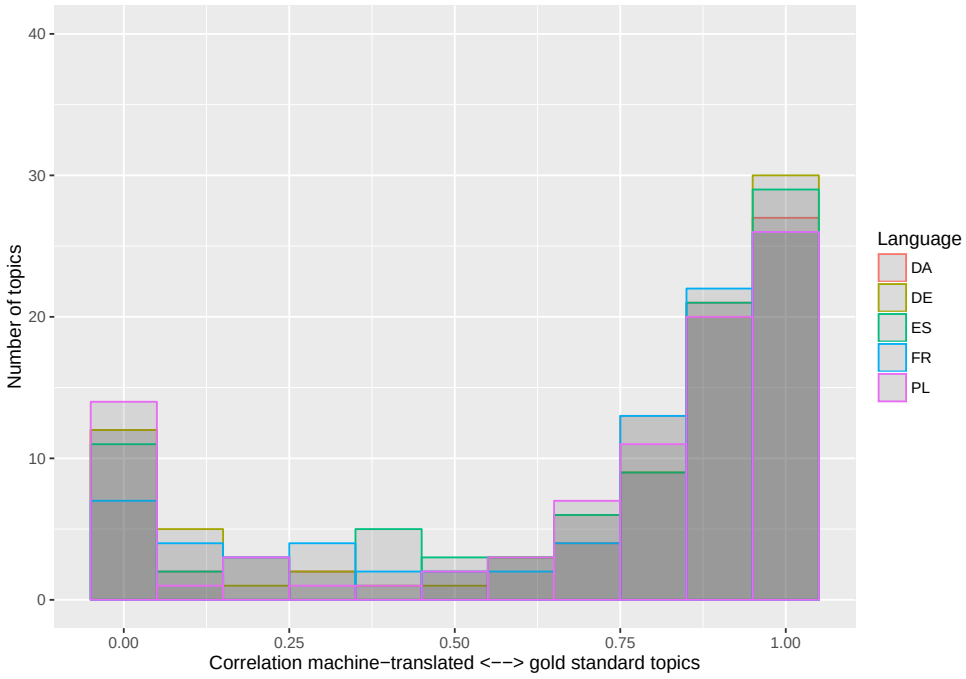
Figure 8: Single topic distribution correlation with unequal no. of topics



Overall descriptives: N=449, M=0.740, SD=0.280

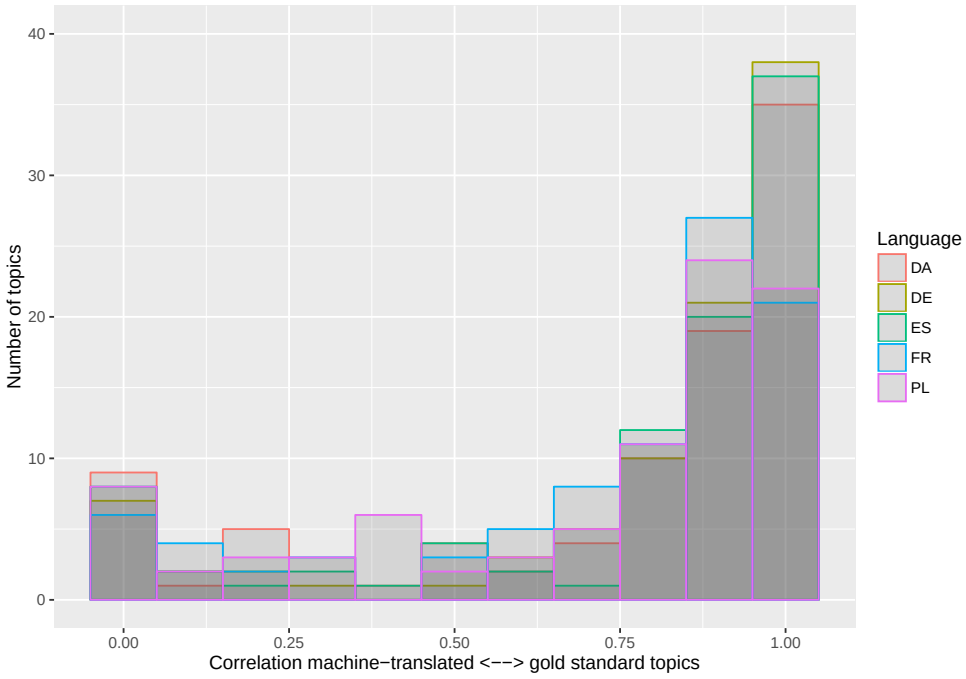
The correlation between topic pairs shows similar results. Here, correlation measures the extent to which the topic scores of a single topic over all documents are similar for the gold standard and machine-translated documents. Figures 7 and 8 show these results for the models with an equal number of topics and an unequal number of topics respectively. The distribution of correlation scores seems to be slightly more skewed to the high end for unequal topic numbers compared to equal topic numbers, but the overall conclusion is again that changing the number of topics does not have a large impact. In contrast to the document-to-document comparison, differences between languages are not statistically significant here.

Figure 9: Topic pair content correlation with equal no. of topics



Overall descriptives: N=446, M=0.708, SD=0.345

Figure 10: Topic pair content correlation with unequal no. of topics



Overall descriptives: N=449, M=0.747, SD=0.315

Finally, in addition to assessing the similarity in topic distributions per document and per topic pair, we also compare the similarity in the actual content of paired topics. To do so, the feature scores of all features that are shared between the gold standard and machine-translated dfm's were compared for each topic pair. The results are presented in figures 9 and 10 for equal and unequal numbers of topics respectively. Again, the distribution of correlation scores per topic pair is more skewed towards the high end for unequal topic numbers compared to equal ones. Also there are no significant differences between the languages.

These results are consistent in that they all show either equal or increased correlation between documents and topics when the gold standard and machine-translated models are run at their respective optimum number of topics, compared to taking the gold standard optimum as overall optimum. This shows that the comparability of topic models does not

decrease and under some circumstances actually increases when they are run at their respective optimums, and as such that running a topic model on machine-translated documents with the optimum number of topics based on that data will result in a highly similar topic model compared to hand-made translations. Also, changing the number of topics does not affect the outcome of the topic model adversely in terms of comparability.

Figure 11: Average proportion of topics with correlation <.7 in documents (equal no. of topics)

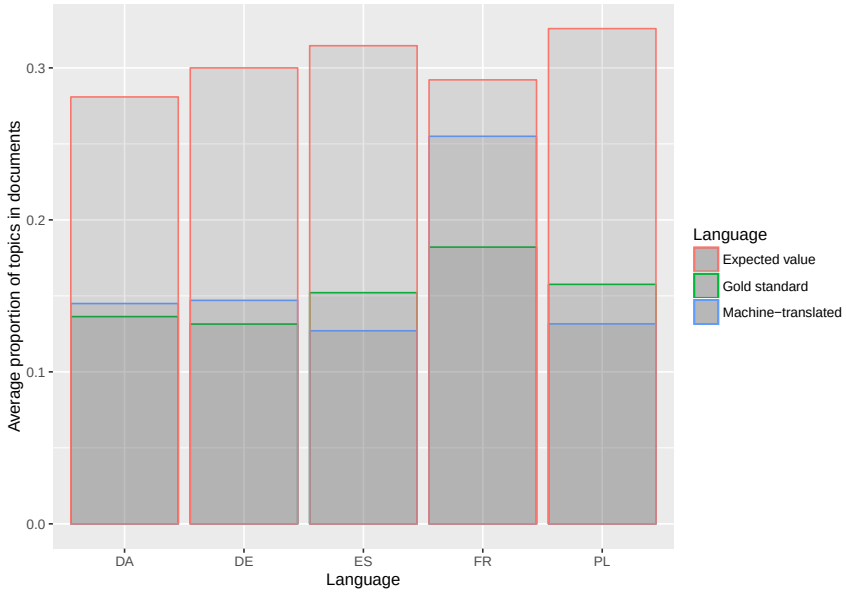
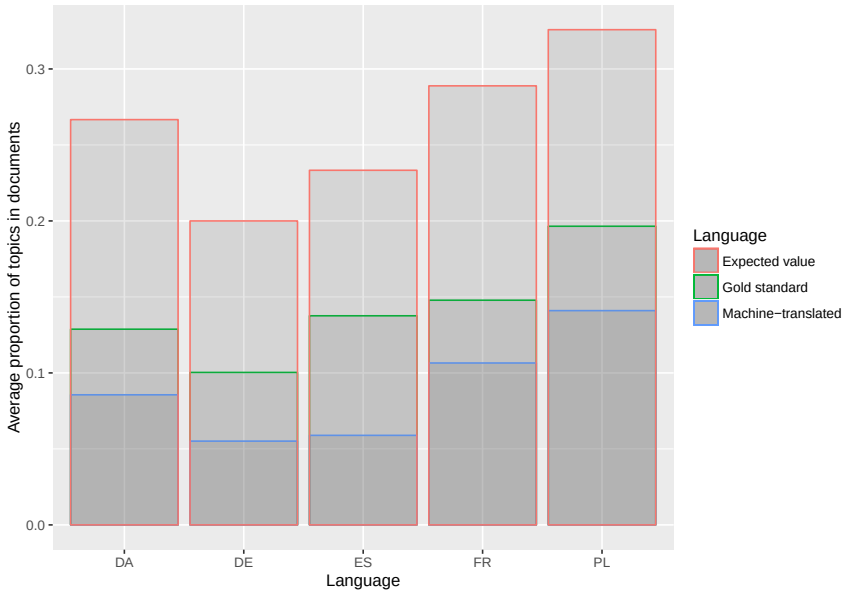


Figure 12: Average proportion of topics with correlation <.7 in documents (unequal no. of topics)



One question that remains to be answered, however, is the spike in topic correlations on the low end of figures 7 through 10, and also why this spike is visible in the topic-to-topic comparisons but not so much in the document-to-document comparisons. The explanation can be found in figures 11 and 12, which show for topic pairs that have a correlation below 0.7 how much these topics are on average present in documents (range 0-1) for both the gold standard and machine-translated models. In addition, the expected proportion of topic pairs with a correlation below 0.7 is also plotted, assuming that all topics have on average an equal share in documents. Again, figure 11 shows this information for models with equal topic numbers, while figure 12 shows it for models with an unequal number of topics.

The most notable difference between the plots for models with an equal and unequal number of topics is that the average expected proportion of these topics in documents is lower with an unequal number of topics instead of an equal one. This is explained by the fact that with different numbers of topics, matches between topics can be more easily made,

as at least 10 of the topics from the machine-translated model are dropped per definition. In addition, it shows that in general the proportion of topic pairs with a correlation below 0.7 decreases. Regardless of the number of topics, the results of both plots show that there is in general a large difference between the observed and expected proportion of these topics in documents, which shows that the topic pairs that have a relatively low correlation, are in general topics that are not very commonly present in documents, and as such not as relevant for the topic models. One result that deviates from this interpretation is the relatively small difference between the observed and expected topic proportions for French machine-translated texts in the comparison of models with an equal number of topics. However, this difference becomes larger, and more in line with the observations for other languages, when looking at the comparison of models with an unequal number of topics. This is also evidence that supports the assumption that when using machine-translated text in topic models, choosing the optimum number of topics based on the actual data is of prime importance.

Conclusion

This paper evaluated the usefulness of machine translation for automated bag-of-words models, comparing bag-of-words vectors of machine translated and gold standard texts. We found only small differences between topic models of hand-translated texts and machine-translated texts. Specifically, across languages we found that over 92% of document pairs achieved a cosine similarity of 0.80 or higher, and although we found significant differences across languages, their effect sizes are small indeed. What is more, we found high overlap in the set of features between hand-translated and machine-translated texts. At the topic level, we found topic distributions to be highly correlated: 65% of document pairs have a correlation of 0.80 or higher. Furthermore, on average 58% (equal number. of topics) or 60% (unequal number of topics) of topic pairs correlated 0.80 or higher in their distribution over all documents.

These findings have important implications for researchers using automated, bag-of-words text analysis methods in their comparative work. In particular, we note that when the amount of text analysts want to compare across languages is large, machine-translation can indeed serve as a useful tool, because on average machine-translated texts appear to be highly similar to hand-translated texts, despite some variation in individual texts. The same goes for the topic distributions within and across texts. In other words, the gold rush may not be over just yet.

References

Agarwal, Apoorv, Boyi Xie, Ilia Vovsha, Owen Rambow and Rebecca Passonneau. 2011. Sentiment analysis of twitter data. In *Proceedings of the workshop on languages in social media*. Association for Computational Linguistics pp. 30–38.

Aharoni, Roee. 2015. Automatic detection of machine translated text and translation quality estimation PhD thesis Department of Computer Science, Bar-Ilan University Ramat Gan, Israel 2015.

Balahur, Alexandra and Marco Turchi. 2014. "Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis." *Computer Speech & Language* 28(1):56–75.

Benoit, Kenneth, Daniel Schwarz and Denise Traber. 2012. The Sincerity of Political Speech in Parliamentary Systems: A Comparison of Ideal Points Scaling Using Legislative Speech and Votes. In *2nd Annual Conference of EPSA, Berlin*. pp. 19–21.

Benoit, Kenneth and Paul Nulty. 2013. "Quanteda: Quantitative Analysis of Textual Data." *An R library for Managing and Analyzing Text* .

Bird, S., E. Klein and E. Loper. 2009. *Natural Language Processing with Python*. O'Reilly Media, Inc.

Blei, David M. 2012. "Probabilistic Topic Models." *Communications of the ACM* 55(4):77–84.

Blei, David M, Andrew Y Ng and Michael I Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3(1):993–1022.

Boumans, Jelle W and Damian Trilling. 2016. "Taking Stock of the Toolkit: An Overview of Relevant Automated Content Analysis Approaches and Techniques for Digital Journalism scholars." *Digital Journalism* 4(1):8–23.

Denny, Matthew James and Arthur Spirling. 2016. "Assessing the Consequences of Text Preprocessing Decisions." *Available at SSRN 2849145* .

Gray, Russell D and Quentin D Atkinson. 2003. "Language-tree divergence times support the Anatolian theory of Indo-European origin." *Nature* 426(6965):435–439.

Greene, Zac, Andrea Ceron, Gijs Schumacher and Z. Fazekas. 2016. "The Nuts and Bolts of Automated Text Analysis. Comparing Different Document Pre-Processing Techniques in Four Countries." *Open Science Framework* (osf.io/4z5z3.).

Grimmer, Justin 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.

Hampshire, Stephen and Carmen Porta Salvia. 2010. "Translation and the Internet: evaluating the quality of free online machine translators." *Quaderns: revista de traducció* (17):197–209.

Hornik, Kurt and Bettina Grün. 2011. "topicmodels: An R package for fitting topic models." *Journal of Statistical Software* 40(13):1–30.

Kaljahi, Zadeh and Rasoul Samad. 2015. The role of syntax and semantics in machine translation and quality estimation of machine-translated user-generated content PhD thesis Dublin City University.

Koehn, Philipp. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*. Vol. 5 pp. 79–86.

Koehn, Philipp and Christof Monz. 2006. Manual and automatic evaluation of machine translation between european languages. In *Proceedings of the Workshop on Statistical Machine Translation*. Association for Computational Linguistics pp. 102–121.

Loaiciga, Sharid, Thomas Meyer and Andrei Popescu-Belis. 2014. English-French Verb Phrase Alignment in Europarl for Tense Translation Modeling. In *LREC*. pp. 674–681.

Lotz, Susan and Alta Van Rensburg. 2014. "Translation technology explored: Has a three-year maturation period done Google Translate any good?" *Stellenbosch Papers in Linguistics Plus* 43:235–259.

Lucas, Christopher, Richard A Nielsen, Margaret E Roberts, Brandon M Stewart, Alex Storer and Dustin Tingley. 2015. "Computer-assisted text analysis for comparative politics." *Political Analysis* p. mpu019.

Popescu-Belis, Andrei, Thomas Meyer, Jeevanthi Liyanapathirana, Bruno Cartoni and Sandrine Zufferey. 2012. Discourse-level annotation over europarl for machine translation: Connectives and pronouns. In *Proceedings of the eighth international conference on Language Resources and Evaluation (LREC)*. Number EPFL-CONF-192582.

Scarton, Carolina and Lucia Specia. 2014. Documentlevel translation quality estimation: exploring discourse and pseudo-references. In *The 17th Annual Conference of the European Association for Machine Translation*. pp. 101–108.