Building a Swedish Question-Answering Model

Hannes von Essen
Chalmers University of Technology
hannes.von.essen@gmail.com

Daniel Hesslow
Chalmers University of Technology
daniel.hesslow@gmail.com

Abstract

High quality datasets for question answering exist in a few languages, but far from all. Producing such datasets for new languages requires extensive manual labour. In this work we look at different methods for using existing datasets to train question-answering models in languages lacking such datasets.

We show that machine translation followed by cross-lingual projection is a viable way to create a full question-answering dataset in a new language. We introduce new methods both for bitext alignment, using optimal transport, and for direct cross-lingual projection, utilizing multilingual BERT.

We show that our methods produce good Swedish question-answering models without any manual work.

Finally, we apply our proposed methods on Spanish and evaluate it on the XQuAD and MLQA benchmarks where we achieve new state-of-the-art values of 80.4 F1 and 62.9 Exact Match (EM) points on the Spanish XQuAD corpus and 70.8 F1 and 53.0 EM on the Spanish MLQA corpus, showing that the technique is readily applicable to other languages.

1 Introduction

The application of supervised machine learning approaches on reading comprehension tasks such as question answering has traditionally been unsuccessful due to a lack of large-scale datasets for training and the lack of powerful enough language models (Hermann et al., 2015). In recent years however, important steps have been made in both areas with the introduction of large-scale datasets (SQuAD (Rajpurkar et al., 2016), SELQA (Jurczyk, Zhai, and Choi, 2016)) and the paradigm-shifting language model BERT.

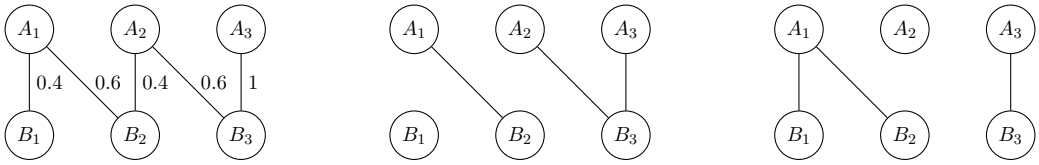
While this has enabled impressive results for English question answering, there is still a lack of such large-scale datasets in other languages such as Swedish. We therefore explore whether it is possible to train a model for Swedish question-answering using the existing English dataset.

Our two main approaches are to (1) fine-tune a multilingual BERT model on the English SQuAD (Stanford Question Answering Dataset) and see how well it generalizes to Swedish, i.e. doing zero-shot learning, and to (2) machine-translate the English dataset into Swedish and fine-tune a Swedish BERT model on it. We also evaluate various combinations of the two. As SQuAD is based on retrieving the answer from a text, the main challenge with the translation to Swedish consists in determining where the beginning and end positions of the answers are in the translated text, i.e. *projecting* the answer span onto the translated sentence. This is difficult as the translation may change the order of the words (e.g. ”The 1973 oil crisis” $\rightarrow$ ”Oljekrisen 1973”), and the translation also changes depending on the context around it.

We experiment with two approaches to the projection problem: one optimal transport-based solution that creates a word-alignment mapping between the English and the Swedish sentence, and one deep learning-based solution that uses multilingual BERT to find the position of the answer given translations of the English answer and a few words surrounding it, with varying amounts of surrounding words included; something we will refer to as a *translation pyramid*.

We evaluate our model for Swedish QA on our machine-translated versions of the SQuAD dev set, but also apply the method on Spanish to evaluate it on two human-made benchmarks, where we establish a new state-of-the-art for Spanish QA systems. We make both the Swedish and Spanish datasets produced by our method freely available[1].

[1] https://github.com/Vottivott/building-a-swedish-qa-model



(a) The translation probabilities between the source and the target text.

(b) The alignment produced by maximizing the probability from A to B.

(c) The alignment produced by maximizing the probability from B to A

Figure 1: A simple example illustrating the difficulty of word-alignment, a reasonable solution is $A_i \overset{\text{Aligned with}}{\longleftrightarrow} B_i$

2 Related work

(Lee et al., 2018) trained a Korean question-answering model using a machine-translated SQuAD in conjunction with a small manually annotated "seed" dataset of Korean question-answer pairs, which is used to predict the translation certainty to remove poor translations from the dataset. Their method for determining the answer span in the translated text is based on modifying the text to be translated by adding citation marks around the answer span. We found this to be unreliable (sometimes citation marks would be removed or shifted in the translation) and to compromise the translation quality. We therefore translate the original text as is, and instead propose a method for reliably finding the translated answer span afterwards.

(Carrino, Costa-jussà, and Fonollosa, 2019) used an automatic method to translate SQuAD into Spanish. They use a custom-trained neural machine translation model to translate the dataset and then create word alignments using the methods described in (Östling and Tiedemann, 2016) to find the answer span in the translated text. This method is similar to our optimal transport method but uses a different way to find the word alignments.

3 Problem definition

The extractive question answering task we want to solve is defined as follows. Given a context paragraph c (for the SQuAD dataset, these are paragraphs from Wikipedia articles) and a question q, we want to extract the answer from the context paragraph c. This consists in predicting the start and the end position of the answer in c, so that our predicted answer becomes $c_{start:end}$.

The SQuAD dataset essentially consists of tuples of $(c, q, start, end)$ and the translation of the dataset requires not only retrieving the text translations c' and q', but also finding the corresponding ground-truth $start'$ and end' such that the answer is given by $c'_{start':end'}$. This is what we refer to as the *projection* problem.

4 Method

4.1 Projection method I: Optimal Transport Based Word Alignment

We introduce a novel optimal transport-based word alignment method. While optimal transport has been used in NLP to describe the distance between different texts (eg. Word Movers Distance, (Kusner et al., 2015)) we are instead interested in the transport plans which roughly describes which words correspond to which.

Optimal transport is used to account for the fact that given word-wise translation probabilities between the source and the target text, the translation of one word is affected by the translation of other words. See figure 1 for an example. The discrete Wasserstein optimal transport problem, (Villani, 2008), is defined as

$$\underset{\gamma}{\operatorname{argmin}} < \gamma, M >_F$$
$$s.t. \quad \gamma \mathbb{1} = p, \quad \gamma^T \mathbb{1} = q, \quad \gamma \geq 0$$

where M is the cost matrix, in the case of word alignment it is the negative log of the pairwise translation probabilities. p is the mass for each word in the source and q is the mass for each word in the target, and γ is a matrix describing a transport plan which conserves the mass.

Additionally any good word alignment method must take the position of the source words and target words into account when finding the alignments. Traditionally this has been done by biasing the translation probabilities towards the diagonal, see for example (Dyer, Chahuneau, and Smith, 2013). However to achieve this we instead look at the discrete Gromov-Wasserstein optimal transport problem, (Mémoli, 2011),

$$\underset{\gamma}{\operatorname{argmin}} \sum_{i,i',j,j'} \left\|d_{i,i'} - \overline{d}_{j,j'}\right\|_2^2 * \gamma_{i,j} * \gamma_{i',j'}$$
$$s.t. \quad \gamma \mathbb{1} = p, \quad \gamma^T \mathbb{1} = q, \quad \gamma \geq 0$$

Here d and $\overline{d}$ describe the pairwise distances between the words in the source and the target sentences respectively. Note that the distance functions may not only take into account the index of the word in the sentence but could for example also be designed such that the distance is larger between words that are in different sentences than between words that are in the same sentence. Intuitively this optimization problem tries to find the transport plan which maintains the pairwise distances before and after the translation (Vayer et al., 2018a). Finally, both of these optimization problems can be combined into the so-called fused Gromov-Wasserstein optimal transport problem first introduced in (Vayer et al., 2018b),

$$\underset{\gamma}{\operatorname{argmin}} (1-\alpha) * < \gamma, M >_F +$$
$$\alpha * \sum_{i,i',j,j'} \left\|d_{i,i'} - \overline{d}_{j,j'}\right\|_2^2 * \gamma_{i,j} * \gamma_{i',j'}$$
$$s.t. \quad \gamma \mathbb{1} = p, \quad \gamma^T \mathbb{1} = q, \quad \gamma \geq 0$$

(Vayer et al., 2018b) also propose an optimization method based on conditional gradients to find a local minimum, we use the open source implementation from (Flamary and Courty, 2017). It is worth noting that to be able to use this method successfully one must be able to estimate the pairwise translation probabilities, the mass of each word as well as the distances within the sentences. While finding sufficiently good approximations for the distances within a sentence is easy it is more difficult to find the masses and translation probabilities for all words. We use the cosine similarity of the supervised multilingual word embeddings provided in (Conneau et al., 2017) as M and set the mass of p and q to be uniform.

4.2 Projection method II: Span Projection using Cross-Lingual BERT

Instead of calculating an alignment between all words in the sentences and extracting the translated answer span from it, a different approach is to try to find the projection of the span directly, without any intermediate step. We propose a method that trains on a cross-lingual (mixed English/Swedish) version of the task we want to solve, and then counts on the generalization abilities of multilingual BERT to be able to apply it on the fully Swedish end task.

More specifically, when training, the task is to predict a span of words in an English sentence given Swedish translations of the span and its surrounding words, while in the application of the method the task is performed on a *Swedish* sentence given the same Swedish translations of the span and its surrounding words. This allows us to use the SQuAD dataset itself for training.

The input to the model is designed to both give information about the span and its surroundings, in order to ensure that the correct position is selected when there are multiple occurrences of the answer text. We use Swedish translations of the following: the answer span, the range from two words before to two words after the answer span, and the range from five words before to five words after the answer span, as illustrated in Figure 2. These are sent as input to the projection model, separated by [SEP] tokens, along with the full English (for training) or Swedish (for the end application) sentence. The output heads are identical to the ones used for the SQuAD task itself, allowing us to reuse most of the code from the SQuAD training.

An additional benefit of the multiple translations is increased robustness due to the variation in the translations it creates. With multiple variations of the translation, there is a greater chance that at least one of them will be more similar to the translated paragraph. For example, there is often ambiguity in whether titles of movies and other works should be translated or left in their original language. With the multiple translations, there is a greater chance that one of them will match the language used in the translated paragraph.

One potential benefit of training using SQuAD rather than a more general dataset could be that it priors the network towards predicting answer-like spans, which could help the network make more plausible guesses in unclear situations.

While we used the answer spans from SQuAD for both the training and the application, it would be possible to generate more training data by using random spans from the text instead of just the answer spans, although this would reduce the potential benefit discussed in the last paragraph.

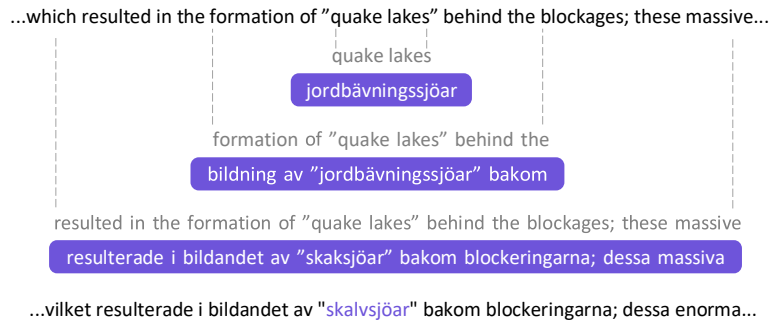


Figure 2: Illustration of the translation pyramid used to give contextual information to the span projection model. All three translations are fed into the network, separated by [SEP] tokens. Note that the answer, *quake lakes*, is translated differently in the expansion levels as either *jordbävningssjöar* or *skaksjöar*, and that both are different from the sentence translation where it is translated as *skalvsjöar*. This illustrates one reason why surrounding words can be important for extra information, and also shows the variation between the different translations, which can make the projection system more robust.

4.3 SQuAD training

We experimented with 4 different language models as the basis for our training: the Base version of Multilingual BERT (cased) (Devlin et al., 2018), the Base and Large version of a Swedish BERT model trained by the Swedish Public Employment Service[2] (uncased) and XLM (Lample and Conneau, 2019) (cased). For the multilingual models we used the pretrained models provided by (Wolf et al., 2019). Their script for SQuAD training also served as the basis for our code.

The multilingual BERT model is trained on Wikipedia in 104 different languages, while the XLM model is trained on Wikipedia in 17 languages. Similarly, the Swedish models are trained on the complete Swedish version of Wikipedia. All models are thus trained on the same amount of Swedish data. It should be noted that the English Wikipedia is larger than the Swedish Wikipedia so we can expect the multilingual models to be slightly better at English than at Swedish.

4.4 Experiments

All experiments were run for 160 000 training steps with a batch size of 3, and an initial learning rate of $5 \cdot 10^{-6}$ that decreases linearly until it reaches zero at step 160 000.

For the deep learning-based projection method, we used an initial learning rate of $3 \cdot 10^{-5}$, a batch size of 8, and we trained with a linearly decreasing learning rate reaching zero after 2 epochs.

For the translation of the paragraphs, questions and answers, as well as the excerpts needed for the translation pyramid, Google Translate's free document translation service[3] was used.

The datasets used are called *en* (the original English SQuAD v1.1 dataset), *exact-sv* (the dataset produced by keeping only the answers in which there is a single exact match of the translated answer in the translated text), *exact-and-ot-sv* (the combination of *exact-sv* with the application of the optimal transport method on the answers removed by *exact-sv*), and *proj-sv* (the application of the deep learning-based projection method on the full dataset). The corresponding versions of the dev set are named similarly as *sv-dev-proj*, *sv-dev-exact* and *en-dev*. In the SQuAD training data there is only one answer per question, while in SQuAD dev each question has a list of multiple acceptable answers, averaging 3.3 per question. Because of the filtering-out of answers with no exact match, most questions in *sv-dev-exact* have fewer answers than in the original dev set, with an average of 1.1 answers per question. Therefore the performance on *sv-dev-exact* is expected to be inherently lower than on *sv-dev-proj*.

5 Results

5.1 Span Projection Methods

While there was no ground-truth dataset to evaluate the performance of the projection models, 200 random examples not used in the training were manually inspected to give statistical estimations

[2]https://github.com/af-ai-center/bert

[3]https://translate.google.com/#view=home&op=docs

of their performance. The results are summarized in Table 1 and 2, where the errors were divided into five categories. For the deep learning-based solution (Projection method II), we found no ***correct answer but wrong position*** errors and only one ***completely incorrect answer*** error, with most errors being ***partially incorrect answer*** errors. This indicates that the method is really good at finding the correct general position of the span in the text. It should also be noted that many of the ***partially incorrect answer*** errors are very small, sometimes missing only a single letter, and can often still be considered good answers to the question. Some examples were also identified as impossible to solve; ***impossible due to translation*** means that the translation has shifted words around such that a correct projection would require the span to be split into multiple parts, making a correct projection using only a single span impossible, and ***impossible due to tokenization*** means that a correct projection would require single tokens to be split into smaller pieces.

Error type	#	Percent
Correct answer but wrong position	0	0%
Partially incorrect answer	10	5%
Completely incorrect answer	1	0.5%
Impossible due to translation	3	1.5%
Impossible due to tokenization	2	1%
Total	**16**	**8%**

Table 1: Evaluation of Projection method II

Error type	#	Percent
Correct answer but wrong position	0	0%
Partially incorrect answer	37	18.5%
Completely incorrect answer	7	3.5%
Impossible due to translation	3	1.5%
Impossible due to tokenization	1	0.5%
Total	**48**	**24%**

Table 2: Evaluation of Projection method I

One issue that causes ***impossible due to tokenization*** is that definiteness is part of the word in Swedish (e.g. "hund**en**") while it is a separate article in English ("**the** dog"). Occasionally the English dataset doesn't include the definite article in the span and the Swedish dataset is not tokenized as to include the indefinite form of the word as a separate token. The model will then sometimes split on the closest boundary causing it to output a result which isn't a word.

The results for the optimal transport method (method I) on the same 200 examples are listed in Table 2, showing that the deep learning-based approach gives considerably more reliable results.

5.2 Evaluation of the Swedish QA models

The results are listed in Table 3. For each metric, we list the highest score achieved across all training checkpoints. We can see that the Multilingual BERT model significantly outperformed the XLM model in our experiments. For Multilingual BERT, the *exact-and-ot-sv* is better than using only *exact-sv*, but *proj-sv* is better than both *exact-and-ot-sv* and *exact-sv*. Additionally, we can see that adding the original English dataset to the training mix gives a small additional improvement in all metrics. Similarly, the performance on *en-dev* for these multilingual mixes is higher than *en*, i.e. the addition of Swedish data to English SQuAD improves the English performance in our experiments. Interestingly the English performance of the models trained only on Swedish data is also high, with *proj-sv* being only 1.8 F1 and 2.2 EM points lower than *en*. Also, the devoted Swedish models perform much worse than multilingual BERT, even when trained on the same Swedish-only data.

5.3 Evaluation on Spanish benchmarks

As there are no benchmarks for Swedish question answering available, we also applied the method on Spanish and evaluated the trained models on the newly introduced XQuAD (Artetxe, Ruder, and Yogatama, 2019) and MLQA (Lewis et al., 2019) benchmarks. XQuAD consists of professional translations of parts of the SQuAD dev set, while MLQA consists of thousands of new QA examples in different languages, crowdsourced from Wikipedia. For the evaluation we used the model checkpoint from the last step of the training. The results, shown in Table 4, show that our method beats the state of the art across all metrics.

6 Conclusions and Future Work

6.1 Conclusions

We have presented a method to automatically translate the question answering datasets with high quality and show that training on such datasets results in good models.

Model	Training data	F1 / EM (sv-dev-proj)	(sv-dev-exact)	(en-dev)
Multilingual BERT Base	en	75.0 / 63.9	69.2 / 57.2	88.8 / 81.5
	exact-sv	76.6 / 64.4	73.6 / 62.8	83.9 / 75.1
	exact-and-ot-sv	80.4 / 69.3	74.5 / 62.3	86.8 / 78.8
	proj-sv	81.4 / 71.5	74.9 / 62.8	87.0 / 79.3
	en + exact-and-ot-sv	80.9 / 70.0	75.1 / 63.0	89.6 / 82.6
	en + proj-sv	**81.9 / 71.6**	**75.6 / 63.3**	**89.8 / 82.8**
XLM	en + proj-sv	74.6 / 64.0	67.8 / 56.0	81.5 / 74.0
Swedish BERT Base	exact-sv	56.8 / 44.0	56.7 / 44.6	
	exact-and-ot-sv	62.6 / 49.3	60.0 / 45.8	
	proj-sv	64.3 / 51.7	60.7 / 47.1	
Swedish BERT Large	exact-sv	56.5 / 43.1	56.7 / 44.6	
	exact-and-ot-sv	62.0 / 48.2	60.0 / 45.8	
	proj-sv	63.9 / 51.4	60.3 / 46.7	

Table 3: Evaluation of the Swedish QA models

Dataset	Model	Training data	F1 / EM
XQUAD	Our models	proj-es	79.8 / 62.1
		en + proj-es	**80.4 / 62.9**
	(Carrino, Costa-jussà, and Fonollosa, 2019)	TAR-train + mBERT (SQuAD-es)	77.6 / 61.8
	XQuAD mBERT baselines	JointMulti 32k voc	59.5 / 41.3
		JointMulti 200k voc	74.3 / 55.3
		JointPair with Joint voc	68.3 / 47.8
		JointPair with Disjoint voc	72.5 / 52.5
MLQA	Our models	proj-es	70.0 / 52.2
		en + proj-es	**70.8 / 53.0**
	(Carrino, Costa-jussà, and Fonollosa, 2019)	TAR-train + mBERT (SQuAD-es)	68.1 / 48.3
	MLQA mBERT baselines	mBERT	64.3 / 46.6
		Translate-train + mBERT	53.9 / 37.4
		XLM (MLM + TLM, 15 languages)	68.0 / 49.8

Table 4: Results for the Spanish evaluation on XQuAD and MLQA

We conclude that limitations in the amount of available Swedish data for pre-training of BERT and the reduced quality of BERT that comes from this can be compensated for to some extent by having larger amounts of data in other languages such as English, which makes multi-lingual models a promising tool for applications in low-resource languages. Even when treated as a Swedish BERT model and ignoring its multi-lingual capacities, the multi-lingual BERT model is probably the best Swedish BERT model currently available, as it outperforms the devoted Swedish model to a remarkable degree (by 17.6 F1 points) in our task when fine-tuned on the same Swedish data. We would therefore recommend Swedish NLP-practitioners to use the multi-lingual BERT model rather than the existing Swedish BERT models until Swedish BERT models trained on larger Swedish datasets become available. We also conclude that when fine-tuning multilingual BERT for an end-task in a certain language (in our case Swedish or Spanish), there can be a benefit in also mixing in training data from other languages than the end-task language .

6.2 Future work

In order to alleviate the problem with poor translations, a future direction could be to try to incorporate a model for quantifying the translation certainty in order to remove the worst translations from the training set, as was proposed by (Lee et al., 2018). Uncertain projections could also be filtered out by looking at the distribution of the output logits from the projection model.

References

Artetxe, M.; Ruder, S.; and Yogatama, D. 2019. On the cross-lingual transferability of monolingual representations. *CoRR* abs/1910.11856.

Carrino, C. P.; Costa-jussà, M. R.; and Fonollosa, J. A. R. 2019. Automatic spanish translation of the squad dataset for multilingual question answering.

Conneau, A.; Lample, G.; Ranzato, M.; Denoyer, L.; and Jégou, H. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dyer, C.; Chahuneau, V.; and Smith, N. A. 2013. A simple, fast, and effective reparameterization of ibm model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 644–648.

Flamary, R., and Courty, N. 2017. Pot python optimal transport library.

Hermann, K. M.; Kociský, T.; Grefenstette, E.; Espeholt, L.; Kay, W.; Suleyman, M.; and Blunsom, P. 2015. Teaching machines to read and comprehend. *CoRR* abs/1506.03340.

Jurczyk, T.; Zhai, M.; and Choi, J. D. 2016. Selqa: A new benchmark for selection-based question answering. *CoRR* abs/1606.08513.

Kusner, M.; Sun, Y.; Kolkin, N.; and Weinberger, K. 2015. From word embeddings to document distances. In *International conference on machine learning*, 957–966.

Lample, G., and Conneau, A. 2019. Cross-lingual language model pretraining. *CoRR* abs/1901.07291.

Lee, K.; Yoon, K.; Park, S.; and Hwang, S.-w. 2018. Semi-supervised training data generation for multilingual question answering. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA).

Lewis, P.; Oğuz, B.; Rinott, R.; Riedel, S.; and Schwenk, H. 2019. Mlqa: Evaluating cross-lingual extractive question answering. *arXiv preprint arXiv:1910.07475*.

Mémoli, F. 2011. Gromov–wasserstein distances and the metric approach to object matching. *Foundations of computational mathematics* 11(4):417–487.

Östling, R., and Tiedemann, J. 2016. Efficient word alignment with markov chain monte carlo. *The Prague Bulletin of Mathematical Linguistics* 106(1):125–146.

Rajpurkar, P.; Zhang, J.; Lopyrev, K.; and Liang, P. 2016. Squad: 100, 000+ questions for machine comprehension of text. *CoRR* abs/1606.05250.

Vayer, T.; Chapel, L.; Flamary, R.; Tavenard, R.; and Courty, N. 2018a. Fused gromov-wasserstein distance for structured objects: theoretical foundations and mathematical properties. *arXiv preprint arXiv:1811.02834*.

Vayer, T.; Chapel, L.; Flamary, R.; Tavenard, R.; and Courty, N. 2018b. Optimal transport for structured data with application on graphs. *arXiv preprint arXiv:1805.09114*.

Villani, C. 2008. *Optimal transport: old and new*, volume 338. Springer Science & Business Media.

Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; and Brew, J. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv* abs/1910.03771.

Appendix A: 10 randomly selected examples from the generated datasets

Who was in control of the Dutch East India Company (VOC) and the Dutch West India Company (WIC)?	**The States General of the United Provinces** were in control of the Dutch East India Company (VOC) and the Dutch West India Company (WIC), but some shipping expeditions were initiated by some of the provinces, mostly Holland and/or Zeeland.
Vem hade kontroll över det nederländska East India Company (VOC) och det Dutch West India Company (WIC)?	**USA: s generalsekreterare** hade kontroll över det nederländska östindiska kompaniet (VOC) och det nederländska västindiska kompaniet (WIC), men vissa sjöfartsekspeditioner initierades av några av provinserna, främst Holland och / eller Zeeland.
¿Quién tenía el control de la Compañía Holandesa de las Indias Orientales (VOC) y la Compañía Holandesa de las Indias Occidentales (WIC)?	**Los Estados Generales de las Provincias Unidas** tenían el control de la Compañía Holandesa de las Indias Orientales (VOC) y la Compañía Holandesa de las Indias Occidentales (WIC), pero algunas de las provincias iniciaron algunas expediciones marítimas, principalmente Holanda y / o Zelanda.

How many pubs applied to be allowed to sell alcohol 24 hours a day?	The Licensing Act 2003, which came into force on 24 November 2005, consolidated the many laws into a single Act. This allowed pubs in England and Wales to apply to the local council for the opening hours of their choice. It was argued that this would end the concentration of violence around 11.30 pm, when people had to leave the pub, making policing easier. In practice, alcohol-related hospital admissions rose following the change in the law, with alcohol involved in 207,800 admissions in 2006/7. Critics claimed that these laws would lead to "24-hour drinking". By the time the law came into effect, 60,326 establishments had applied for longer hours and **1,121** had applied for a licence to sell alcohol 24 hours a day. However nine months later many pubs had not changed their hours, although some stayed open longer at the weekend, but rarely beyond 1:00 am.
Hur många pubar ansökte om att få sälja alkohol 24 timmar om dygnet?	Licenslagen 2003, som trädde i kraft den 24 november 2005, konsoliderade de många lagarna till en enda lag. Detta gjorde det möjligt för pubar i England och Wales att ansöka till kommunfullmäktige för de öppettider som de valde. Det hävdades att detta skulle avbryta koncentrationen av våld omkring klockan 11.30, när människor var tvungna att lämna puben, vilket underlättar polisarbetet. I praktiken ökade alkoholrelaterade sjukhusinläggningar efter lagändringen, med alkohol involverad i 207 800 inläggningar under 2006/7. Kritiker hävdade att dessa lagar skulle leda till "24-timmars dricka". När lagen trädde i kraft hade 60 326 anläggningar ansökt om längre timmar och **1 112** hade ansökt om tillstånd att sälja alkohol 24 timmar om dygnet. Emellertid nio månader senare hade många pubar inte ändrat sina timmar, även om vissa stannade öppna längre på helgen, men sällan efter kl.
¿Cuántos pubs solicitaron que se les permita vender alcohol las 24 horas del día?	La Ley de Licencias de 2003, que entró en vigor el 24 de noviembre de 2005, consolidó las numerosas leyes en una sola Ley. Esto permitió que los pubs en Inglaterra y Gales se postularan ante el consejo local para el horario de apertura de su elección. Se argumentó que esto terminaría con la concentración de violencia alrededor de las 11.30 p. M., Cuando la gente tenía que abandonar el pub, lo que facilitaba la vigilancia. En la práctica, los ingresos hospitalarios relacionados con el alcohol aumentaron después del cambio en la ley, con alcohol involucrado en 207,800 ingresos en 2006/7. Los críticos afirmaron que estas leyes conducirían a "beber 24 horas". Cuando entró en vigencia la ley, 60,326 establecimientos habían solicitado más horas y **1,121** habían solicitado una licencia para vender alcohol las 24 horas del día. Sin embargo, nueve meses después, muchos pubs no habían cambiado sus horarios, aunque algunos permanecían abiertos más tiempo el fin de semana, pero rara vez más allá de la 1:00 a.m.

What was the European Union tasked with managing?	Italy became a major industrialized country again, due to its post-war economic miracle. The European Union (EU) involved the division of powers, with taxation, health and education handled by the nation states, while the EU had charge of **market rules, competition, legal standards and environmentalism**. The Soviet economic and political system collapsed, leading to the end of communism in the satellite countries in 1989, and the dissolution of the Soviet Union itself in 1991. As a consequence, Europe's integration deepened, the continent became depolarised, and the European Union expanded to subsequently include many of the formerly communist European countries – Romania and Bulgaria (2007) and Croatia (2013).
Vad fick EU att hantera?	Italien blev igen ett stort industrialiserat land på grund av dess ekonomiska mirakel efter kriget. Europeiska unionen (EU) involverade maktfördelningen, med beskattning, hälsa och utbildning som hanterades av nationalstaterna, medan EU hade ansvaret för **marknadsregler, konkurrens, juridiska standarder och miljöhänsyn**. Det sovjetiska ekonomiska och politiska systemet kollapsade, vilket ledde till slutet av kommunismen i satellitländerna 1989, och Sovjetunionens upplösning 1991. Som en följd av detta fördjupades Europas integration, kontinenten depolariserades och Europeiska unionen utvidgades att därefter inkludera många av de tidigare kommunistiska europeiska länderna - Rumänien och Bulgarien (2007) och Kroatien (2013).
¿Qué se encargó de gestionar la Unión Europea?	Italia se convirtió nuevamente en un importante país industrializado, debido a su milagro económico de posguerra. La Unión Europea (UE) implicó la división de poderes, con impuestos, salud y educación manejados por los estados nacionales, mientras que la UE tenía a su cargo las **reglas del mercado, la competencia, los estándares legales y el ambientalismo**. El sistema económico y político soviético se derrumbó, lo que condujo al fin del comunismo en los países satélites en 1989, y la disolución de la propia Unión Soviética en 1991. Como consecuencia, la integración de Europa se profundizó, el continente se despolarizó y la Unión Europea se expandió. para incluir posteriormente a muchos de los países europeos anteriormente comunistas: Rumania y Bulgaria (2007) y Croacia (2013).

How long was Beyonce depressed?	LeToya Luckett and Roberson became unhappy with Mathew's managing of the band and eventually were replaced by Farrah Franklin and Michelle Williams. Beyoncé experienced depression following the split with Luckett and Roberson after being publicly blamed by the media, critics, and blogs for its cause. Her long-standing boyfriend left her at this time. The depression was so severe it lasted for **a couple of years**, during which she occasionally kept herself in her bedroom for days and refused to eat anything. Beyoncé stated that she struggled to speak about her depression because Destiny's Child had just won their first Grammy Award and she feared no one would take her seriously. Beyoncé would later speak of her mother as the person who helped her fight it. Franklin was dismissed, leaving just Beyoncé, Rowland, and Williams.
Hur länge var Beyonce deprimerad?	LeToya Luckett och Roberson blev missnöjda med Mathews hantering av bandet och ersattes så småningom av Farrah Franklin och Michelle Williams. Beyoncé upplevde depression efter splittringen med Luckett och Roberson efter att ha blivit offentligt klandrad av media, kritiker och bloggar för dess sak. Hennes mångaåriga pojkvän lämnade henne just nu. Depressionen var så allvarlig att den varade i **ett par år**, under vilken hon ibland höll sig i sitt sovrum i flera dagar och vägrade att äta något. Beyoncé uttalade att hon kämpade för att tala om sin depression eftersom Destiny's Child just hade vunnit sitt första Grammy-pris och att hon fruktade att ingen skulle ta henne på allvar. Beyoncé skulle senare tala om sin mamma som personen som hjälpte henne att bekämpa den. Franklin avskedades och lämnade bara Beyoncé, Rowland och Williams.
¿Cuánto tiempo estuvo deprimido Beyonce?	LeToya Luckett y Roberson se sintieron descontentos con el manejo de la banda por parte de Mathew y finalmente fueron reemplazados por Farrah Franklin y Michelle Williams. Beyoncé experimentó depresión luego de la separación con Luckett y Roberson luego de ser acusado públicamente por los medios, críticos y blogs por su causa. Su novio de larga data la dejó en este momento. La depresión fue tan severa que duró **un par de años**, durante los cuales ocasionalmente se mantuvo en su habitación durante días y se negó a comer nada. Beyoncé declaró que le costaba hablar sobre su depresión porque Destiny's Child acababa de ganar su primer premio Grammy y temía que nadie la tomara en serio. Beyoncé luego hablaría de su madre como la persona que la ayudó a combatirlo. Franklin fue despedido, dejando solo a Beyoncé, Rowland y Williams.

What company office did Bell make the original cross country phone call from?	In January 1915, Bell made the first ceremonial transcontinental telephone call. Calling from the **AT&T** head office at 15 Dey Street in New York City, Bell was heard by Thomas Watson at 333 Grant Avenue in San Francisco. The New York Times reported:
Vilket företagskontor gjorde Bell det ursprungliga telefonsamtalet från andra länder?	I januari 1915 ringde Bell det första ceremoniella transkontinentala telefonsamtalet. Call från **AT&T** huvudkontor på 15 Dey Street i New York City hördes av Thomas Watson på 333 Grant Avenue i San Francisco. New York Times rapporterade:
¿Desde qué oficina de la compañía hizo Bell la llamada telefónica original?	En enero de 1915, Bell realizó la primera llamada telefónica ceremonial transcontinental. Llamando desde la oficina central de **AT&T** en 15 Dey Street en la ciudad de Nueva York, Thomas Watson escuchó a Bell en 333 Grant Avenue en San Francisco. El New York Times informó:

What was the name of West's fashion line for women?	On October 1, 2011, Kanye West premiered his women's fashion label, **DW Kanye West** at Paris Fashion Week. He received support from DSquared2 duo Dean and Dan Caten, Olivier Theyskens, Jeremy Scott, Azzedine Alaïa, and the Olsen twins, who were also in attendance during his show. His debut fashion show received mixed-to-negative reviews, ranging from reserved observations by Style.com to excoriating commentary by The Wall Street Journal, The New York Times, the International Herald Tribune, Elleuk.com, The Daily Telegraph, Harper's Bazaar and many others. On March 6, 2012, West premiered a second fashion line at Paris Fashion Week. The line's reception was markedly improved from the previous presentation, with a number of critics heralding West for his "much improved" sophomore effort.
Vad hette Wests modelinje för kvinnor?	Den 1 oktober 2011 hade Kanye West premiär för sin dametikett, **DW Kanye West**, vid Paris Fashion Week. Han fick stöd från DSquared2-duon Dean och Dan Caten, Olivier Theyskens, Jeremy Scott, Azzedine Alaïa och Olsen-tvillingarna, som också var närvarande under hans show. Hans debutmodeshow fick blandade till negativa recensioner, allt från reserverade observationer från Style.com till uttalande kommentarer från The Wall Street Journal, The New York Times, International Herald Tribune, Elleuk.com, The Daily Telegraph, Harper's Bazaar och många andra. Den 6 mars 2012 hade West premiär för en andra modelinje vid Paris Fashion Week. Linjens mottagning förbättrades markant från den föregående presentationen, med ett antal kritiker som vädrade West för hans "mycket förbättrade" andra ansträngning.
¿Cómo se llamaba la línea de moda de West para mujer?	El 1 de octubre de 2011, Kanye West estrenó su marca de moda femenina, **DW Kanye West** en la Semana de la Moda de París. Recibió el apoyo del dúo de DSquared2 Dean y Dan Caten, Olivier Theyskens, Jeremy Scott, Azzedine Alaïa y los gemelos Olsen, que también estuvieron presentes durante su show. Su desfile de modas debut recibió críticas mixtas y negativas, desde observaciones reservadas por Style.com hasta comentarios fascinantes de The Wall Street Journal, The New York Times, International Herald Tribune, Elleuk.com, The Daily Telegraph, Harper's Bazaar y muchos otros. El 6 de marzo de 2012, West estrenó una segunda línea de moda en la Semana de la Moda de París. La recepción de la línea mejoró notablemente de la presentación anterior, con una serie de críticos que anunciaron a West por su "mucho mejor" esfuerzo de segundo año.

Along with cabaret, striptease, bands and drama, what is a type of stage performance that can be found in pubs?	A few pubs have stage performances such as serious drama, **stand-up comedy**, musical bands, cabaret or striptease; however juke boxes, karaoke and other forms of pre-recorded music have otherwise replaced the musical tradition of a piano or guitar and singing.[citation needed]
Tillsammans med kabaret, striptease, band och drama, vad är en typ av scenuppträdande som finns på pubar?	Några pubar har scenuppträdanden som seriöst drama, **stand-up komedi**, musikband, kabaret eller striptease; men jukeboxar, karaoke och andra former av förinspelad musik har annars ersatt den musikaliska traditionen för ett piano eller gitarr och sång.
Junto con el cabaret, el striptease, las bandas y el drama, ¿cuál es un tipo de actuación en el escenario que se puede encontrar en los pubs?	Algunos pubs tienen representaciones teatrales como dramas serios, **comedias**, bandas musicales, cabaret o striptease; sin embargo, los juke boxes, el karaoke y otras formas de música pregrabada han reemplazado la tradición musical de un piano o guitarra y canto. [cita requerida]

The structure of Bern's city centre is mainly what type of buildings?	The structure of Bern's city centre is largely **medieval** and has been recognised by UNESCO as a Cultural World Heritage Site. Perhaps its most famous sight is the Zytglogge (Bernese German for "Time Bell"), an elaborate medieval clock tower with moving puppets. It also has an impressive 15th century Gothic cathedral, the Münster, and a 15th-century town hall. Thanks to 6 kilometres (4 miles) of arcades, the old town boasts one of the longest covered shopping promenades in Europe.
Strukturen i Berns centrum är främst vilken typ av byggnader?	Strukturen i Berns centrum är till stor del **medeltida** och har erkänts av UNESCO som ett kulturellt världsarv. Det kanske mest kända synet är Zytglogge (Bernese tyska för "Time Bell"), ett genomtänkt medeltida klocktorn med rörliga dockor. Den har också en imponerande gotisk domkyrka från 1500-talet, Münster och ett rådhus från 1500-talet. Tack vare 6 kilometer arkader, har gamla stan en av de längsta täckta shoppingpromenaderna i Europa.
¿La estructura del centro de la ciudad de Berna es principalmente qué tipo de edificios?	La estructura del centro de la ciudad de Berna es en gran parte **medieval** y ha sido reconocida por la UNESCO como Patrimonio Cultural de la Humanidad. Quizás su vista más famosa es el Zytglogge (alemán bernés para "Time Bell"), una elaborada torre de reloj medieval con marionetas en movimiento. También tiene una impresionante catedral gótica del siglo XV, el Münster, y un ayuntamiento del siglo XV. Gracias a 6 kilómetros (4 millas) de arcadas, el casco antiguo cuenta con uno de los paseos comerciales cubiertos más largos de Europa.

In what city are the New York Red Bulls based?	In soccer, New York City is represented by New York City FC of Major League Soccer, who play their home games at Yankee Stadium. The New York Red Bulls play their home games at Red Bull Arena in nearby **Harrison, New Jersey**. Historically, the city is known for the New York Cosmos, the highly successful former professional soccer team which was the American home of Pelé, one of the world's most famous soccer players. A new version of the New York Cosmos was formed in 2010, and began play in the second division North American Soccer League in 2013. The Cosmos play their home games at James M. Shuart Stadium on the campus of Hofstra University, just outside the New York City limits in Hempstead, New York.
I vilken stad är New York Red Bulls baserade?	I fotboll representeras New York City av New York City i Major League Soccer, som spelar sina hemmamatcher på Yankee Stadium. New York Red Bulls spelar sina hemmamatcher på Red Bull Arena i närheten av **Harrison, New Jersey**. Historiskt sett är staden känd för New York Cosmos, det mycket framgångsrika tidigare professionella fotbollslaget som var det amerikanska hemmet Pelé, en av världens mest kända fotbollsspelare. En ny version av New York Cosmos bildades 2010 och började spela i den andra divisionen North American Soccer League 2013. Cosmos spelade sina hemmamatcher på James M. Shuart Stadium på campus vid Hofstra University, precis utanför New York City gränser i Hempstead, New York.
¿En qué ciudad se encuentran los Red Bulls de Nueva York?	En el fútbol, la ciudad de Nueva York está representada por el New York City FC de Major League Soccer, que juega sus partidos en casa en el Yankee Stadium. Los Red Bulls de Nueva York juegan sus partidos en casa en el Red Bull Arena en la cercana **Harrison, Nueva Jersey**. Históricamente, la ciudad es conocida por el Cosmos de Nueva York, el exitoso ex equipo de fútbol profesional que fue el hogar estadounidense de Pelé, uno de los jugadores de fútbol más famosos del mundo. Una nueva versión del New York Cosmos se formó en 2010, y comenzó a jugar en la segunda división de la Liga de Fútbol de América del Norte en 2013. El Cosmos juega sus partidos en casa en el estadio James M. Shuart en el campus de la Universidad de Hofstra, a las afueras de New Límites de la ciudad de York en Hempstead, Nueva York.

How many people came to visit New York in 2013?	Tourism is a vital industry for New York City, which has witnessed a growing combined volume of international and domestic tourists – receiving approximately 51 million tourists in 2011, **54 million** in 2013, and a record 56.4 million in 2014. Tourism generated an all-time high US$61.3 billion in overall economic impact for New York City in 2014.
Hur många besökte New York 2013?	Turism är en viktig industri för New York City, som har sett en växande kombinerad mängd internationella och inhemska turister - med cirka 51 miljoner turister under 2011, **54 miljoner** 2013 och rekord 56,4 miljoner 2014. Turismen genererade en tid höga 61,3 miljarder US dollar i total ekonomisk påverkan för New York City 2014.
¿Cuántas personas vinieron a visitar Nueva York en 2013?	El turismo es una industria vital para la ciudad de Nueva York, que ha sido testigo de un creciente volumen combinado de turistas internacionales y nacionales: recibió aproximadamente 51 millones de turistas en 2011, **54 millones** en 2013 y un récord de 56.4 millones en 2014. El turismo generó un récord histórico alto impacto económico general de US $ 61.3 mil millones para la ciudad de Nueva York en 2014.

Appendix B: Examples of span projection errors

Partially incorrect answer (5%)

In what decade was seafloor spreading discovered?	In the **1960s**, a series of discoveries, the most important of which was seafloor spreading, showed that the Earth's lithosphere, which includes the crust and rigid uppermost portion of the upper mantle, is separated into a number of tectonic plates that move across the plastically deforming, solid, upper mantle, which is called the asthenosphere. There is an intimate coupling between the movement of the plates on the surface and the convection of the mantle: oceanic plate motions and mantle convection currents always move in the same direction, because the oceanic lithosphere is the rigid upper thermal boundary layer of the convecting mantle. This coupling between rigid plates moving on the surface of the Earth and the convecting mantle is called plate tectonics.
Under vilket decennium upptäcktes havsbotten spridning?	På **1960**-talet visade en serie upptäckter, vars viktigaste var havsbotten spridning, att jordens litosfär, som inkluderar jordskorpan och den styva översta delen av den övre manteln, är uppdelad i ett antal tektoniska plattor som rör sig över det plastiska deformerande, fast, övre mantel, som kallas asthenosfären. Det finns en intim koppling mellan plattans rörelse på ytan och konvektionen av manteln: oceaniska plattrörelser och mantelkonvektionsströmmar rör sig alltid i samma riktning, eftersom den oceaniska litosfären är det styva övre termiska gränsskiktet i konvektionsmanteln . Denna koppling mellan styva plattor som rör sig på jordens yta och konvektionsmanteln kallas plattaktonik.

Completely incorrect answer (0.5%)

Why is Warsaw's flora very rich in species?	The flora of the city may be considered very rich in species. The species richness is mainly due to the **location** of Warsaw within the border region of several big floral regions comprising substantial proportions of close-to-wilderness areas (natural forests, wetlands along the Vistula) as well as arable land, meadows and forests. Bielany Forest, located within the borders of Warsaw, is the remaining part of the Masovian Primeval Forest. Bielany Forest nature reserve is connected with Kampinos Forest. It is home to rich fauna and flora. Within the forest there are three cycling and walking trails. Other big forest area is Kabaty Forest by the southern city border. Warsaw has also two botanic gardens: by the Łazienki park (a didactic-research unit of the University of Warsaw) as well as by the Park of Culture and Rest in Powsin (a unit of the Polish Academy of Science).
Varför är Warszawas flora mycket rik på arter?	Stadens flora kan anses vara mycket rik på arter. Artrikligheten beror främst på **Warszawa**s läge inom gränsområdet för flera stora blommoregioner som omfattar betydande andelar nära vildmarksområden (naturskogar, våtmarker längs Vistula) samt åkermark, ängar och skogar. Bielany Forest, som ligger inom gränserna till Warszawa, är den återstående delen av den masoviska urskogen. Naturreservatet Bielany Forest är anslutet till Kampinos Forest. Det är hem till rik fauna och flora. Inom skogen finns tre cykel- och vandringsleder. Ett annat stort skogsområde är Kabaty Forest vid den södra stadsgränsen. Warszawa har också två botaniska trädgårdar: av Łazienki-parken (en didaktisk-forskningsenhet vid universitetet i Warszawa) samt av parken för kultur och vila i Powsin (en enhet av den polska vetenskapsakademin).

Impossible due to translation (1.5%)

Note how the first word of the English answer, "attacked" (in Swedish "attackerade"), has been moved many words back in the Swedish translation, making it impossible to include it without also including many words that are not in the English answer.

During withdrawal from Fort William Henry, what did some Indian allies of French do?	French irregular forces (Canadian scouts and Indians) harassed Fort William Henry throughout the first half of 1757. In January they ambushed British rangers near Ticonderoga. In February they launched a daring raid against the position across the frozen Lake George, destroying storehouses and buildings outside the main fortification. In early August, Montcalm and 7,000 troops besieged the fort, which capitulated with an agreement to withdraw under parole. When the withdrawal began, some of Montcalm's Indian allies, angered at the lost opportunity for loot, **attacked the British column**, killing and capturing several hundred men, women, children, and slaves. The aftermath of the siege may have contributed to the transmission of smallpox into remote Indian populations; as some Indians were reported to have traveled from beyond the Mississippi to participate in the campaign and returned afterward having been exposed to European carriers.
Vad gjorde några indiska allierade franska under utträdet från Fort William Henry?	Franska oregelbundna styrkor (kanadensiska speider och indier) trakasserade Fort William Henry under första halvan av 1757. I januari övergick de bakhåll mot brittiska räknare nära Ticonderoga. I februari inledde de en vågig raid mot positionen över den frusna sjön George och förstörde lagerhus och byggnader utanför huvudbefästningen. I början av augusti beleirade Montcalm och 7000 trupper fortet, som kapitulerade med ett avtal om att dra sig tillbaka under parol. När tillbakadragandet började, attackerade några av Montcalms indiska allierade, arga över den förlorade möjligheten till **plundring, den brittiska kolumnen** och dödade och fångade flera hundra män, kvinnor, barn och slavar. Efterdyningarna av belägringen kan ha bidragit till överföringen av smittkoppor till avlägsna indiska populationer; som vissa indier rapporterades ha rest från utanför Mississippi för att delta i kampanjen och återvända efter att ha blivit utsatta för europeiska transportörer.

Impossible due to tokenization (1%)

Note that the original answer is incorrectly missing the final digit of the act, while the Swedish version has added it back. Even though it improves the final dataset, it is nevertheless an error in the context of cross-lingual projection. Since 1855 is treated as a single token we consider this "impossible due to tokenization".

What document formed the Parliament of Victoria?	Victoria has a written constitution enacted in 1975, but based on the 1855 colonial constitution, passed by the United Kingdom Parliament as the **Victoria Constitution Act 185**5, which establishes the Parliament as the state's law-making body for matters coming under state responsibility. The Victorian Constitution can be amended by the Parliament of Victoria, except for certain "entrenched" provisions that require either an absolute majority in both houses, a three-fifths majority in both houses, or the approval of the Victorian people in a referendum, depending on the provision.
Vilket dokument bildade Victoria parlamentet?	Victoria har en skriftlig konstitution som antogs 1975, men baserad på den koloniala konstitutionen 1855, som antogs av Storbritanniens parlament som **Victoria Constitution Act 1855**, som fastställer parlamentet som statens lagstiftande organ för frågor som kommer under statligt ansvar. Den viktorianska konstitutionen kan ändras av Victoria parlament, med undantag för vissa "förankrade" bestämmelser som kräver antingen en absolut majoritet i båda husen, en tre femtedelars majoritet i båda husen eller godkännande av det viktorianska folket i en folkomröstning, beroende på bestämmelsen.