The Effect of Negative Sampling Strategy on Capturing Semantic Similarity in Document Embeddings

Marzieh Saeidi
marzieh@gluru.co

Ritwik Kulkarni
ritwik@gluru.co

Theodosia Togia
sia@gluru.co

Michele Sama
michele@gluru.co

Abstract

In many machine learning tasks, a model needs to be presented with both correct and incorrect examples during the training. For instance, given a query, a search engine can be trained to predict the relevant (positive) documents from irrelevant (negative) ones. While each query is associated with a handful of relevant documents, the number of irrelevant documents can be vast. This imbalance can bias a document retrieval model while the mere volume of irrelevant documents can result in long training times. In this paper, we show the affect of a tailored negative sampling on the performance of the Deep Structure Semantic Model (DSSM). We show that a naive random sampling method outperforms more sophisticated ways of selecting negative data.

1 Introduction

In many machine learning tasks, it is often the case that negative examples vastly outnumber positive ones. For instance, in a recommendation system, the number of items that a user has rated or purchased are far less than the total number of items available in the system. Similarly, in an answer retrieval task for the community question answering platforms, the number of relevant answers to a question is very low compared to the total number of available answers. To address this imbalance, negative examples are typically down-sampled in accordance with some criteria. The effect of negative sampling strategy has been studied across a range of machine learning and NLP tasks, with varying results (Wunsch et al., 2009; Ustalov et al., 2017; Bordes et al., 2014).

Sampling from the space of negative examples can be performed on the basis of three alternative intuitions, namely that negative examples should be: **i)** as similar as possible to positive examples to increase a model's discriminative power within small neighbourhoods **ii)** as different as possible to positive examples to avoid providing a model with conflicting information **iii)** representative of the entire space of negative examples, without preference for those similar or dissimilar to positive ones

The first intuition has led to *local sampling* methods (e.g. Wunsch et al. 2009), that is sampling from negative cases that are close to positive ones by some specified proximity measure, which can be either linguistically justified or based on vector space distances. The reason is to select or generate data that challenges a model's capacity to distinguish between seemingly similar items. The second intuition is associated with *distance sampling* (Wunsch et al., 2009), namely selecting negative cases that are as distinct as possible from positive ones in order to ensure the data are properly clustered in the vector space. The final intuition has led to *uniform sampling* methods, that is randomly sampling from the space of negative examples.

Adopting the correct negative sampling approach is largely task-dependent (see Section 5). Our research sought to determine the most suitable strategy for learning semantic similarity. In these particular experiments we focused on the widely adopted DSSM model (Huang et al., 2013) trained on the task of answer selection in the community question answering platform Yahoo! Answers.

Q1: *Help im scared! Dental problems?* two of my top adult teeth are a little bit loose, will i lose them? **A1:** If they are adult teeth, you dentist may be able to "tighten" them ...
Q2: *Do I need 2 dental bridges or just one huge one on my teeth?* I have 2 maryland bridges on my top teeth (the teeth on either side of my two top ... **A2:** I would suggest cantilever bridges will be more than enough ...
Q3: *Can Someone Give Me Links Proving Global Warming Real Or Not?* I believe global warming is real, but lately I realized I don't actually know that global warming is real ... **A3:** There are many reason why global warming can occur. From increased solar input to changes in atmospheric aerosol concentration to albedo and land use changes

Table 1: The first two QA pairs are under the category and subcategory *Health-Dental* category. The last QA is under *Environment-Global Warming*.

2 Model

We perform answer selection using the DSSM model Huang et al. (2013). The DSSM is a popular feedforward neural network for document retrieval.[1] It can be seen as a supervised version of latent semantic models (e.g. Latent Semantic Indexing Deerwester et al. (1990)), leveraging implicit document-query relevance judgments, originally from clickthrough data. The model non-linearly projects queries and documents to a reduced dimensionality semantic space which better captures dependencies between the dimensions of the original vector space. It does so by using an objective function which maximises the log probability of the relevant document (i.e. correct answer in the case of question answering) being the most similar. The input of DSSM is one query, one relevant document (positive example) and a number of irrelevant documents (negative examples), each one encoded as a bag of a character trigrams feature vector, as in the original paper.

2.1 Negative Sampling

In the dataset we used for these experiments, Yahoo! Answers, each question and answer group (QA) contains a question, a list of answers with one often flagged as the best answer. Additionally, each QA typically has a specified category and subcategory. In the experiments, given a question, we train the DSSM to maximise its ability to rank a relevant answer higher than any non-relevant one. This leads to a model whose representations of questions and answers belonging to the same topic are in the same neighbourhood. Ideally, a trained model should be not only capable of distinguishing *between* documents in different topics (i.e. to cluster question or answer vectors by their subject matter) but also *within* topics (i.e. to arrange vectors inside each topic in a way that reflects similarity of information need) , with similar questions and their relevant answers forming a sub-cluster within a topic cluster).

A straightforward way to improve the model's ability to discriminate across topics is to explicitly provide this information as input to the model during the training. However, increasing the model's discriminative power inside same-topic (i.e. category) neighbourhoods without compromising per subject clustering is more challenging. Table 1 shows examples of question from two different categories: Health-Dental and Environment- Global Warming. As we can see, QAs within the same category are more likely to have common and similar terms.

As one can assume, it may be easier for a model to distinguish among documents from different categories than those within the same category. Addressing this challenge might require a principled approach to negative sampling, that is deciding which positive and negative examples to provide the model with. For instance, if the model is fed with a politics-related question, what will be the most informative negative answers?

[1] We refer readers to this paper for more modelling details.

We can choose between random sampling, local sampling (e.g. providing negative examples from the same topic), distance sampling (e.g. providing negative answers outside the same topic, which have more pronounced differences with the correct response), or a combination of both.

In the experiments that follow, we compare the effect of two types of negative sampling on the performance of DSSM when topic information is provided to the model:

- Random sampling whereby negative examples are sampled uniformly from the space of all answers
- Refined sampling which is explained in details in Section 2.1.

In refined negative sampling, the model is presented with questions within the same category and subcategory (local sampling) as well as other categories (distance sampling) with a view to training a model capable of generating embeddings that are appropriately clustered not only across different categories, but also within each category. We use category and subcategory information available for each question in order to sample negative examples during the training. We sample from answers that belong to the same category, from answers that belong to the same subcategory and from answers that do not belong to the category of the question. We choose these proportions heuristically with 50% from inside the category (in particular, 20% from the subcategory and 30% from the entire category) and the remaining 50% from outside the category.

Figure 1 shows the number of QAs in each of the 26 categories in our dataset of 3 million QAs. As we can see, the number of QAs across different categories is not balanced. Using random sampling, we may never come across questions from the same category or subcategory for a given QA. This issue can be alleviated using refined sampling.

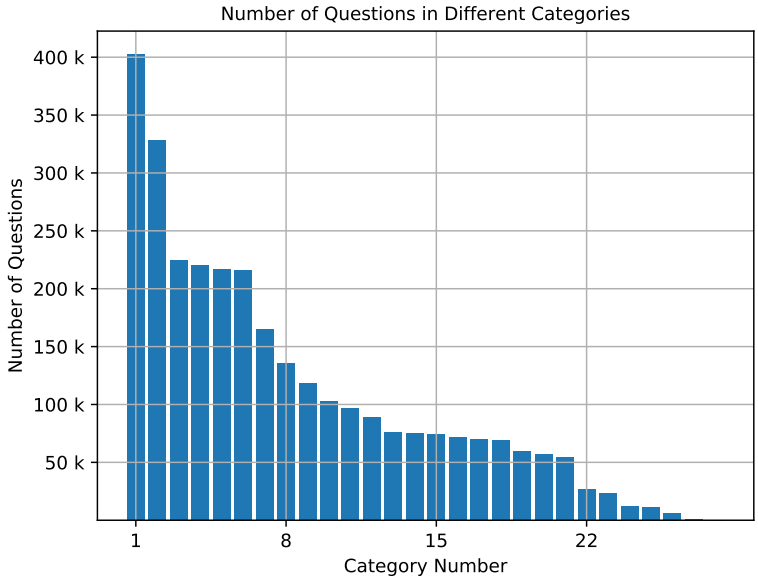

Figure 1: Number of QAs belonging to different 26 categories. Total number of QAs in our dataset is 3 million.

3 Experiments

We run our experiments on a corpus of Yahoo! Answers questions and answers, which contains over three million questions and answers.[2] Topic information is available by means of categories and subcategories logged by users when submitting a question. In our experiments, we use questions for which the

[2] https://webscope.sandbox.yahoo.com

following information is available: **i)** the 'best answer' chosen by the user (used as the positive example), **ii)** the category and subcategory label. Negative examples are chosen among other answers according to the preferred negative sampling strategy.

The authors of the DSSM paper (Huang et al., 2013) mention that they do not observe any significant difference by using different sampling strategies for the unclicked (irrelevant) documents. However, they do not discuss the negative sampling strategies. In this paper, we study the effect of a refined negative sampling strategy using the category information for the task of answer retrieval in the community question answering platforms. Intuitively, *random* sampling has limitations because random examples might not represent all the nuances of semantic dissimilarities.

We hypothesize that providing informative negative examples can help the model to learn more separable embeddings across different topics and within the same topic. Moreover, refined negative sampling may help the model to converge faster (since we provide it with information we believe is more vital from early on). We examine this hypothesis using the three experiments described below.

No Category In this experiment, even though we have access to the category information, in order to provide a baseline, no category information is provided during the training.[3] Since our refined negative sampling strategy uses category information, we only employ random negative sampling.

Category - Random Negative Sampling In this experiment, category and subcategory information is included in the representation of each document as one-hot vectors. Negative answers for each question are sampled randomly from the pool of all answers across all categories and subcategories.

Category - Refined Negative Sampling Similarly to the above experiment, here the category and subcategory information is included. However, instead of random negative sampling, we use the refined negative sampling method described in Section 2.1.

At test time, when searching for the correct answer, we search through two spaces. First, since the user has specified the category of the question, we use this category information to limit the search space by filtering for answers within the same category. This experiment tests the extent to which embeddings are separable within the same category. Second, we search through all the answers across all the categories. This is because we want to test whether the learned embeddings of questions are separable from the answers both globally and in the neighbourhood of the same category.

Evaluation We divide our dataset into train, dev and test sets. We use the dev set to decide when to stop the training and test set to report the evaluation results. While we run the experiments with a varying number of QAs in the training set, the dev and the test set each contain 20k QAs. We report results using Mean Recall@1, Mean Recall@5 and MRR (mean reciprocal rank) (Voorhees et al., 1999).

4 Results

In this section, we look at the performance of the model when trained **i**) with random negative sampling without the category information, tested on the two search spaces described above (*no_ctg* search within category and *no_ctg_all* for global search), **ii**) with random negative sampling and category information, tested on the same two search spaces (*ctg_rnd* and *ctg_rnd_all*) and using refined negative sampling with category information (*ctg_rfn* and *ctg_rfn_all*).

Figure 2 shows the performance of the models, in terms of Mean Recall@1. On the left hand side, the performance is illustrated as the number of negative samples during the training increases. This is done for the case when we search through the entire space (dashed line) and when we search through the category only (solid line). As expected, providing the model with category information increases retrieval performance. Among the models that use categories, the model trained with random negative sampling (blue line) outperforms the one trained with the refined negative sampling (green line). When searching only within the same category, the model with the random negative sampling performs considerably higher than the model trained using refined negative sampling. Moreover, unlike the model with refined

[3]It has been shown that adding category information improves the accuracy of finding the relevant answer in community question answering (Zhang et al., 2016). We also expect that including category information, will result in a better accuracy.

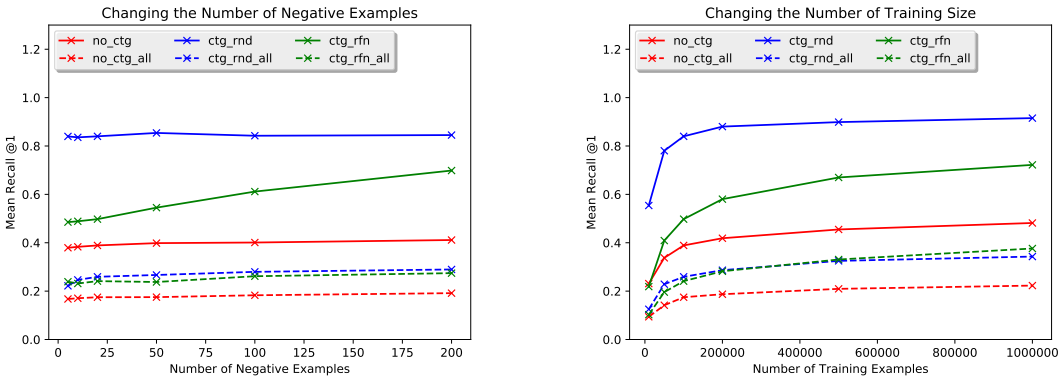

Figure 2: Performances of the models (in terms of *Mean Recall@1*) over the test set as the size of negative examples increases, with training size = 100k (top) and as the size of training examples increases, with negative size = 100 (bottom).

negative sampling, that does not reach a plateau, the model with the random negative sampling converges to the optimum performance without requiring a large number of negative examples during training.On the other hand, the performance of the model with the refined negative sampling does not reach a plateau even with 200 negative examples. On the right hand side of the figure, we see the performance of the models in terms of Mean Recall@1, as the size of training data increases. Similar to the previous figure, we can see that by adding category information (blue and green lines), search performance increases considerably. We see that the refined negative sampling has a negative effect on the performance of the DSSM. Further, the model with the random negative sampling reaches optimum performance using only 200k samples whereas the model that uses refined negative sampling does not reach a plateau even with 1m samples. Performance of the models in terms of Mean Recall@5 and MRR are illustrated in Figures 3 and 4. Similar patterns can be seen on these figures as well.

Further, by observing the dev loss during the training (Figure 5 in the Appendix), we can see that learning is stable across the epochs in the case of random sampling given the category information. However, when sampling is refined, there is a considerable amount of oscillation, which can reflect the model's effort to resolve the conflict between items in the same category being similar in some ways and different in others.

Overall, we observe that the refined negative sampling has a negative effect on the performance of the model, which is in line with the observation in Huang et al. (2013) for the document retrieval task.

The reason why refined sampling underperforms random sampling might be that DSSM mainly relies on text similarity in order to find relevant documents or answers. This is especially the case when working with longer documents such as QAs (as opposed to the experiments the authors of DSSm Huang et al. (2013) carried on document titles only) By providing negative examples from the same category, the model will see similar text patterns as negative examples (in the example in Table 1 the word "dental" has appeared in two questions within same category). This can possibly also explain why the loss fluctuates during training epochs when refined sampling is applied: the model is given conflicting information as to whether similar character trigrams are relevant to each other. However, as the number of negative examples or training examples increases, the model adjusts its weights so that similar terms in two documents make the documents relevant to each other.

4.1 Analysis

Figure 2 empirically demonstrates that the refined negative sampling results in performance loss compared to random sampling, when category information is provided. To analyse this behaviour more, we examine how the choice of negative sampling strategy (i.e. random or refined) affects the formation of vector space neighbourhoods whose vectors must share the same subject (e.g. 'Entertainment'). In par-

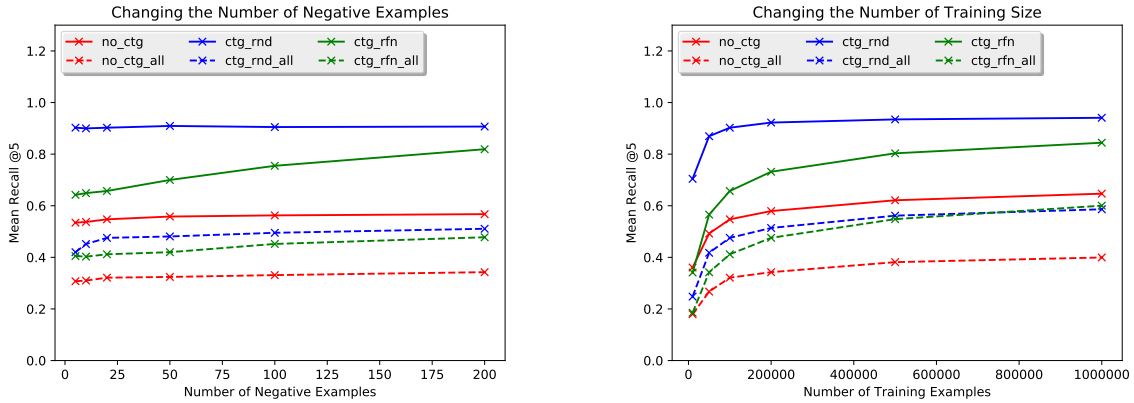

Figure 3: Performances of the models (*Mean Recall@5*) over the test set as the size of negative examples increases, training size = 100k (left) and as the size of training examples increases, negative size = 100 (right).

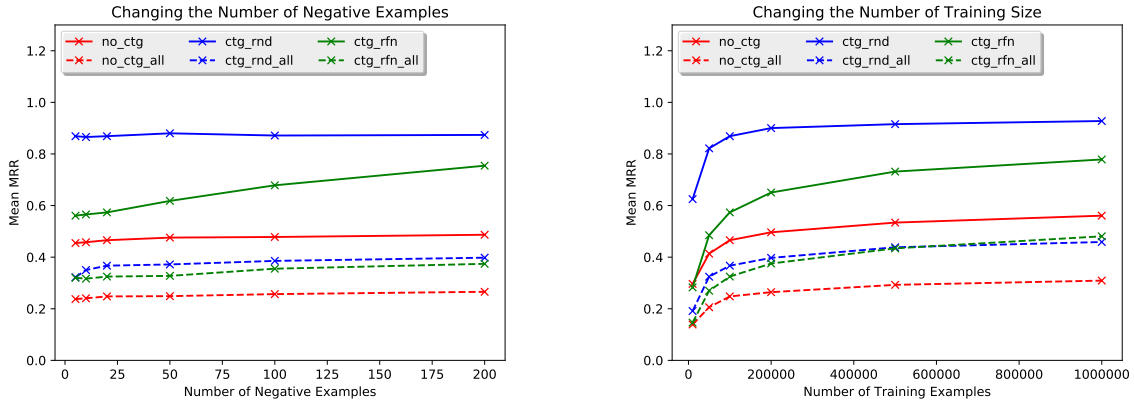

Figure 4: Performances of the models (*MRR*) over the test set as the size of negative examples increases, training size = 100k (left) and as the size of training examples increases, negative size = 100 (right).

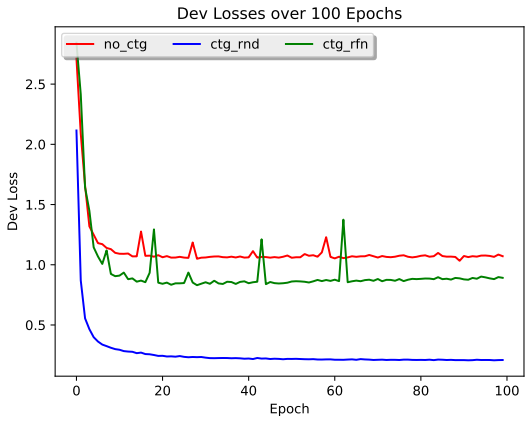

Figure 5: Dev loss during the training epochs.

ticular, we compare the two sampling approaches as to how similar their formed neighbourhoods are to the ideal clusters, which result from grouping vectors by their original Yahoo! Answers categories.

We start by generating embeddings using the DSSM for the two negative sampling strategies that include category information during training, namely, "Category with random negative sampling" ((*ctg_rnd* and *ctg_rnd_all* and "Category with refined negative sampling" (*ctg_rfn* and *ctg_rfn_all*). To explore the extent to which category information is encoded in the embeddings, we cluster the embeddings for each of the two cases independently and compare them with the ideal clusters, that are obtained by collecting together embeddings that have the same category. We perform k-means clustering with *k* set to the number of unique categories in the Yahoo! Answers dataset.

The two sets of clusters (one for each negative sampling strategy) are compared with the ideal Yahoo! Answers clusters using Mutual Information. If the clustering of embeddings agrees to a high degree with the Yahoo! Answers clusters then we can conclude that the category labels find a direct representation in the 128 dimensional embedding space, where the category information is among the principal components along which the questions are arranged. Mutual information, which here is used to quantify the dependence between each of the two sets of clusters with the ideal set (i.e. Yahoo! Answers clusters), can be calculated as follows:

$$I(X, Y) = H(X) + H(Y) - H(X; Y)$$

where $I(X, Y)$ is the mutual information between sets X and Y, $H(X)$ is the entropy of the set X, while $H(X; Y)$ is the joint entropy of the sets X and Y.

	YC-Emb_R	YC-Emb_N
I	0.492	0.365
I/H(YC)	0.118	0.087

Table 2: Mutual information between Yahoo! Answers clusters and embedding clusters with refined negative sampling ($YC - Emb\_N$) or embedding clusters with random negative sampling ($YC - Emb\_R$)

The very low mutual information between the embeddings and Yahoo clusters for both with and without refined negative sampling (Table 2) indicates that the category labels do not have a significant influence on how the embeddings are arranged in the vector space. We can observe that the clusters formed when DSSM is trained with random negative sampling (Emb_R) have higher mutual information with the ideal Yahoo! Answers clusters compared to that of refined negative sampling clusters (Emb_N) and Yahoo.This adds to the notion developed from figures 2, 3 and 4 that refined negative sampling disrupts the learning across and within category examples.

5 Related work

Mikolov et al. Mikolov et al. (2013) introduce "negative sampling" as a way of approximating the calculation of softmax over the entire vocabulary for each forward step. The effect of negative sampling strategy has been studied across a range of NLP tasks, with varying results. Wunsch et al. (2009) perform local sampling for anaphora resolution by selecting non-anaphoric pairs from sentences surrounding anaphoric pairs (i.e. positive data). In another experiment, they use as negative examples the ones misclassified by a classifier trained on positive data only. However, they obtain their best results using random sampling. Similarly, for the task of hypernymy extraction, Ustalov et al. (2017) boost the performance of their system by selecting a false hypernym for a word from the neighbourhood of the correct hypernym. They report a drop in performance when random sampling is applied. Bordes et al. (2014) obtain negative entity-relation-entity triples for open question answering by partially corrupting positive ones, creating negative examples that are more challenging for a model to discriminate. Logacheva et al. (2015) improve their machine translation system by generating negative examples from variations of the positive ones. For the task of co-reference resolution, Zhekova (2011) achieves improvements

by adjusting the ratio of positive to negative examples. Negative sampling has also been investigated in the context of discourse relation extraction (Li and Nenkova, 2014) and classification of caused motion constructions (Hwang and Palmer, 2015).

6 Conclusions

In this paper, we examine the impact of a refined negative sampling on the performance of the Deep Structured Semantic Model (Huang et al., 2013) when applied to an answer retrieval task in community question answering. We found that random sampling, albeit less sophisticated, is a more reliable method for providing negative examples to DSSM, as had originally been observed for the task of web search. As a future experiment, the same negative sampling techniques can be used with a model other than DSSM. This can shed light on whether this effect is due to the model's inherent limitations in learning semantic information, which restrict its ability to handle more subtle negative examples or whether it is a more general trend in the task of answer retrieval for community question answering.

References

Bordes, A., J. Weston, and N. Usunier (2014). Open Question Answering with Weakly Supervised Embedding Models. *Ecml*, 165–180.

Deerwester, S., S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman (1990). Indexing by latent semantic analysis. *Journal of the American society for information science 41*(6), 391.

Huang, P.-s., N. M. A. Urbana, X. He, J. Gao, L. Deng, A. Acero, and L. Heck (2013). Learning Deep Structured Semantic Models for Web Search using Clickthrough Data. *the 22nd ACM international conference on Conference on information & knowledge management*, 2333–2338.

Hwang, D. J. and M. Palmer (2015). Identification of Caused Motion Construction. *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*, 51–60.

Li, J. J. and A. Nenkova (2014). Addressing Class Imbalance for Improved Recognition of Implicit Discourse Relations. *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)* (June), 142–150.

Logacheva, V. and L. Specia (2015). The role of artificially generated negative data for quality estimation of machine translation. *18th Annual Conference of the European Association for Machine Translation*, 51 – 58.

Mikolov, T., K. Chen, G. Corrado, and J. Dean (2013). Distributed Representations of Words and Phrases and their Compositionality. *Nips*, 1–9.

Ustalov, D., N. Arefyev, C. Biemann, and A. Panchenko (2017). Negative Sampling Improves Hypernymy Extraction Based on Projection Learning. *2*, 543–550.

Voorhees, E. M. et al. (1999). The trec-8 question answering track report. In *Trec*, Volume 99, pp. 77–82.

Wunsch, H., S. Kübler, and R. Cantrell (2009). Instance Sampling Methods for Pronoun Resolution. *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, 478–483.

Zhang, K., W. Wu, F. Wang, M. Zhou, and Z. Li (2016). Learning distributed representations of data in community question answering for question retrieval. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, pp. 533–542. ACM.

Zhekova, D. (2011). Zhekova - Instance Sampling for Multilingual Coreference Resolution. *7*(September), 150–155.