The Resolution of Gender Anaphora Reference with the Help of Kernel Trick Mechanisms

A A Pak, S Amirzhan, A A Ziyaden
Institute of Informational and Computational Technologies, Almaty, Kazakhstan
aa.pak83@gmail.com, amir.sanzhar@gmail.com, iamdenay@gmail.com

Abstract. An article represents the results of the experiments constructing of conventional anaphora resolution system for gender-balanced corpora GAP. The main idea behind the conducted set of experiments is taking advantage over the kernel trick mechanism of deep word vectorization for various semantic and syntax features. We will show that it has a positive impact on system performance versus the traditional approach.

1. Introduction
Coreference Resolution (CR) is a crucial step in the pipeline of many high-level applications of Natural Language Processing (NLP), namely fact extraction, machine translation, text summarization, etc. The task of CR begins its history from the first attempts to resolve the connection between pronoun and noun phrases [1], nowadays, such a task called Anaphora Resolution (AR). Subsequently, at Message Understanding Conference, some partial subproblems [2] were resolved. Traditionally, algorithms, as the first step, generated the set of mentions of entities. After that, each pair of pronouns and mention was classified with the help of a decision tree algorithm. At the end of processing the positively classified pairs were merged into a cluster, that represents itself the dedicated chain between connected contexts of mentions and noun phrases. In the paper [3] the modification of generative grammar was presented, namely «Government and Binding theory» (GB). The significant part of the theory was dedicated to syntax anaphora, i.e., analysis of links between pronoun and other parts of the speech.

The next epoch of coreference resolution begins with the conference "ConLL-2011/2012" and the commonly accessed task of modeling non-limiting coreferences; as a result, a new variety of approaches appeared [4]. Coreference is a link between mentions (noun phrases), which are related to the same entity. By analogy with GB, there is an approach that utilized multi filtration, where each level of filter setups or denies a link between words. At the paper [5] it was indicated that on the general noun phrase coreference task the learning approach achieves accuracy comparable to that of nonlinear approaches, moreover, instead of considering the task of AR as a local optimization problem, authors proposed new consideration of the task as global optimization among all pairs in the document.

There are a wide variety of manually marked resources and corpora. Nonetheless, most of them do not capture ambiguous pronouns in enough volume or diversity to accurately indicate the practical applicability of models. Furthermore, there is a lack of balanced corpora due to gender bias in existing corpora and systems favoring masculine entities. A new corpus was presented in order to overcome these shortcomings, a gender-balanced labeled corpus (GAP) of 8,908 ambiguous pronoun-name pairs sampled to provide diverse coverage of challenges posed by the real-world text. There is a range of baselines that demonstrates the complexity of the challenge; the best result is just a 66.9% F1 score [6].
2. Word embeddings: BERT
Word embeddings are a form of computational distributional semantics for determining a word's meaning "from the company it keeps" [7,8], i.e., the words it co-occurs. The most popular model is recently presented Bidirectional Encoder Representations from Transformers (BERT). Unlike other word embeddings [9,10] BERT is designed to pre-train deep bidirectional representations by jointly conditioning on both left and right context in all layers. As a result, the mentioned model suits well for a wide range of NLP tasks, such as question answering and coreference resolution and others [11].
BERT's model has several common features with ELMO and OpenAI GPT. It is a multi-layer bidirectional Transformer encoder based on the original realization described in [12]. Nowadays, the implementation of Transformers is so popular that we will omit the exhaustive model architecture description and refer readers to Vaswani et al. [12].

3. Kernel trick
A kernel trick is a mathematical tool that can be applied to any algorithm which consists of a dot product between two vectors. The main idea of the kernel trick is getting a better discriminability of samples' classes with the help of feature space extension. Kernel trick assumes that to compute the dot products of vectors in the higher dimensional space; a kernel function can be used directly with the lower-dimensional vectors. Therefore, every dot product can be replaced by a kernel function. A Kernel function has its area of definition in the original space and returns the dot product of the vectors in the enriched feature space, which can be produced with the help of simple algebraic operations over the original space with low dimension. The feature space is usually a higher dimensional space where data can be linearly separable. In order for a function to be a valid Kernel function, it should fulfill the "Mercer's theorem". According to the Mercers' theorem, every positive definite symmetric function can be seen as a kernel function.

4. Dataset codebook and extra features
The task is to identify the target of a pronoun within a text passage. The source text is taken from Wikipedia articles. In the dataset, there are labels of the pronoun and two candidate names to which the pronoun could refer. An algorithm should be capable of deciding whether the pronoun refers to name A, name B, or neither. The dataset is also available on the GAP Github Repo.
There are the following columns for analysis:
- ID - Unique identifier for an example (Matches to Id in output file format);
- Text - Text containing the ambiguous pronoun and two candidate names (about a paragraph in length);
- Pronoun - The target pronoun (text);
- A - The first name candidate (text);
- B - The second name candidate;
- URL - The URL of the source Wikipedia page for the example;
- Extra features

\[ \tilde{w}_{ab} = \tilde{a} \cdot \tilde{b} \]  \hspace{1cm} (1)

\[ \tilde{w}_{pp} = \tilde{p} \cdot \tilde{p} \]  \hspace{1cm} (2)

\[ \tilde{w} = \exp(\| \tilde{w}_{pp} - \tilde{w}_{ab} \|) \]  \hspace{1cm} (3)
The following features increased the performance of the model: distances between A and B candidates versus pronoun; the flag of the determinant (his or her) usually links to a subject, syntax roles; kernel function as described above. There are 4000 samples with 6149-dimension size.

5. Models and Experiments
Our neural network is Feed-Forward Network (FFN) which consist 2 hidden layers of size 2048 with dropout 0.7 and 254 with dropout 0.2, activation function used 'relu'. Last layer is softmax - probabilistic distribution of classes, see figure 1.

![Figure 1. The architecture of anaphora resolution classifier.](image)

6. Results
The task is to identify the target of a pronoun within a text passage. Together with various vectorization architectures, various additional features were tested. You can see the comparative analysis in Table 1, in the first column, the model name and additional features are given to improve the classification. The second column shows the error in the form of the cross-entropy value.

| Model                              | Validation Loss |
|------------------------------------|-----------------|
| Base BERT                          | 0.63            |
| Large BERT                         | 0.58            |
| Large BERT + features              | 0.52            |
| Large BERT + kernel trick + features| 0.45            |
| Large BERT + [18,20] layers + features| 0.42            |
| Large BERT + ALL                   | 0.33            |

7. Discussion and Conclusion
Labeled corpora of texts for training and testing modern systems of natural language processing have had a statistical bias towards the masculine. In the present work, we considered the classical problem of resolving anaphoric connections as exemplified by the gender-balanced corpus of texts. The problem can be described as finding the chains of interrelated words in meaning from the stream of tokens, the syntax of noun groups and pronouns. It is worth noting the significant advantage of vectorization using BERT compared to traditional methods. In order to improve the model, we tested several additional features, namely the distance between noun groups and a candidate-pronoun in
tokens; a sign of a linguistic determinant; other syntactic features; and the kernel function, which made the most significant contribution to the improvement of the model. A potential direction of development is testing new word vectorization models and expanding the corpus of gender-balanced texts.

Acknowledgements
We gratefully acknowledge financial support of Ministry of Education and Sciences, Republic of Kazakhstan (Grant AP05132760).

References
[1] Hobbs J R 1978 Resolving pronoun references Lingua, 44 (4), pp. 311-338
[2] Seventh Message Understanding Conference: Proceedings of a Conference Held in Fairfax, Virginia, USA, MUC 1998, April 29 - May 1, 1998. ACL 1998
[3] Chomsky N 1981 Lectures in Government and Binding. (Studies in generative grammar 9.) Dordrecht: Foris
[4] Pradhan S, Moschitti A, Xue N, Uryupina O, Zhang Y 2012 CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In Proceedings of the Sixteenth Conference on Computational Natural Language Learning (CoNLL 2012), Jeju
[5] Soon W M, Ng H Y, Lim D C Y 2001 A machine learning approach to coreference resolution of noun phrases. Computational Linguistic(special Issue on Computational Anaphora Resolution),27(4):521-544
[6] Webster K, Recasens M, Axelrod V, Baldridge J 2018 Mind the GAP: A Balanced Corpus of Gendered Ambiguous Pronouns.
[7] Firth J R 1957 Applications of General Linguistics. Transactions of the Philological Society, 56: 1-14. doi:10.1111/j.1467-968X.1957.tb00568.x
[8] Amirzhan S, Pak A, Jaxylykova A B 2018 The estimation of stability of semantic space generated by word embedding algorithms. 1-5. 10.1109/iSAI-NLP.2018.8692814.
[9] Peters M, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, Zettlemoyer L 2018 Deep contextualized word representations. In NAACL
[10] Radford A, Narasimhan V, Salimans T, Sutskever I 2018 Improving language understanding with unsupervised learning. Technical report, OpenAI
[11] Devlin J, Chang M-W, Lee K, Toutanova K 2018 Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805
[12] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser L, Polosukhin I 2017 Attention is all you need. CoRR, abs/1706.03762