A simple but potentially powerful approach for multilingual parsing

Pablo Botton da Costa, Helena de Medeiros Caseli, and Fabio Natanael Kepler

Laboratrio de Lingustica e Inteligncia Computacional (LALIC),
13565-905 So carlos, Brazil
{pablo.costa,helenacaseli}@dc.ufscar.br
{fabiokepler}@unipampa.edu.com

Abstract. All approaches today for multilingual dependency parsing don't use any support of sister languages. This paper present a very different approach to deal with data sparsity, language transfer, etc. Our approach really use sister languages theory, combining resources from that type of languages, we want a model who can really parse many different languages. In this paper, we present a new architecture who not use any type of oracles to make good parsing decisions. We hope that type of approach can really make dynamical parsing decisions.

Keywords: Multilingual, dependency parsing, sister languages, deep learning, language transfer

1 Introduction

Nowadays, several papers [5, 1] demonstrates the importance of combine multilingual resources, and how impactful this combination can be, when compared with monolingual techniques for several different benchmarks for syntactic and semantic means. In particular, [5] demonstrate how important is the language transfer through vector combination from different languages increase the quality of representations, in short the experiments shown the importance of combine languages like english and german are better then combine english with french. This experiments was tests in different benchmarks proposed by [10].

The state of art for language transfer for any task in Natural Language Processing area (NLP), has been suffering an real "revolution" with the increased use of deep learning methods. An great chunk of this "revolution" it's because of the recent efforts in ways how to get better representations of words. In that sense, we quote the excellent works of [10], [16] e [3], which are based in the excellent phrase of (Firth, J. R., 1957) "You shall know a word by the company it keeps".

Some of the studies methods of distributional representation, as [10] and specially and [3], revolutionized the way of how think in NLP. Particularly, for the use of embeddings for the representations of words. This approach for representation, allowed an best handling and representation of the words, and allowed the use of an power-full and forgotten group of algorithms: the neural methods.

The different methods of transfer language, that this paper will address are only methods who combine resources of different languages for better representation of words, or which can combine resources from different languages in order to increase them. The approach presented in this paper will utilize neural methods for distributional representations of words. For that, neural networks are an excellent way of represent words, since they can easily capture the characteristics of words and his context automatically.

Our application here will be in dependency parsing. The syntactic analysis of one sentence consists in generate the proper syntactic representation of an sentence in dependency or constituent form. Our option for the dependency form, is because that type of representation have the capability of easily represent the sentence, and was highly recommended before for multilingual parsing, for many authors as [12] and [15]. This choice is due in short for the simplicity in the representation of its structure and free word order, and that problem are just one of the most important in multilingual syntactic analysis.

Recently, the methods of language transfer has been largely used for dependency parsing. As shown in [14], [9], for many languages poor resource that previously obtained underperforming results compared to languages with high resources, such as English. The use of combined resources can increase the performance of the parser for that cases where the language have a lack of resources. This type of techniques use lexical dictionaries, word cluster[1] and resource transference, in order to obtain more resources or better models to then realize the parsing.

Despite the growing use of transfer language methods for tasks such as parsing [14, 9], as well discuss in [12, 15], none of the proposals submitted until today proposes the linguistic transfer of resources from sister languages.

Thus, at this work is considered as sister languages those that: (i)having greater amount of similarities when compared with other languages, (ii) are present in the same language branch and (iii) have the same mother language in common. For example, the German and Portuguese languages are not sisters, but they can be considered similar languages to be present in the same linguistic branch, the Indo-European.

Only [15] propose the use of sister languages as the logic solution to deal with the problem of morphological richness languages, and poor resource languages. The effort of our work here is measure how important is the utilization of resource from sister languages in language transfer techniques for multilingual parsing. This choice is through the strong argument of [15]: sisters languages would have greater similarity and, in hypothetically would make more sense to be combined.

For the words, in portuguese ”conhecimento” and in spanish ”conocimiento”, these words have identical meaning which, besides having a very similar writing, also are used at same contexts. This similarity is not only phonetical and orthographical, but extends at morphological and syntactical levels as well. The objective of this paper is explore this syntactical/morphological similarity of sis-

[1] Is a techniques for induction of cluster in unsupervised fashion for the lexical extraction, distances between clusters and other major features.

ter languages, to combine them thought language transfer methods to obtain better results in multilingual parsing in order to compare with very well know methods like [9, 14, 7].

In the same manner as [9, 14, 7], our proposal seeks investigate the importance of language transfer, from sister languages, as solution to the inherent problem of poor resource language and morphological richness languages. The focus application here is dependency parsing which, how demonstrates [12] and [15], It reduces the complexity associated with the large number of possible morphological/syntactical symbols presents in that type of language, and data sparsity as well.

Our neural model proposed here to dependency parsing, is completely different from that proposed by [2]. Since our model takes into account all decision possibilities, instead of use an oracle to select the best decision, and his label.

2 The problem

According with [15] and [12], the use of sister languages methods for dependency parsing emerges as a good alternative to reducing a lot of the problems with morphological richness languages (MRL), such as data sparsity, free-order-word.

[2] propose the use of a multilayer perceptron as an decider, it means, base transition decisions in neural choices. This model follows the principle proposed by [6] which use a decider to select which transition is the best option based on a configuration. This configuration it is an pair of word and his respective morphological tags.

An transition-based parser based on arc-standart transitions aims to predict a transition from a initial configuration. This configuration is a set of elements in the buffer, stack and dependency sets. The elements in the buffer are the sentence to parse, The element in the stack is that you want to analyze, and the dependency set is the whole operations the parser already made for that sentence.

In [7] an idea for language transfer is presented based on distributed representations, but without any quote to the sister languages hypothesis. The next sub-section it aims to demonstrate the differences between our proposal and [7]. More specifically, our article presents a novel neural model for dependence analysis based on transition decisions in a multilingual fashion, using language transfer based on sister languages.

2.1 Language transfer

One of the best ways to deal with the problem of poor resource languages, and morphological richness languages is the language transfer. the idea behind is using techniques which can deal with different domains. For example in parsing, the most normal approach are lexical dictionaries. Or, train an model in one language, and using word cluster techniques to deal with different vocabularies, same as [8]. Today, an very power-full technique is train the parser in richness resource language like english, [14] and [9], and parse an poor resource language creating a new "gold corpus", and then re-train the model with that new corpus.

But, very recently [7] proposed to use an dependency parser combined with an correlation techniques, which combine vectors in different languages to predict sentences in an multilingual way. His approach uses of statistical correlation between two sets of vectors, to project them and get a new set of combined vectors, this same approach was very well used before by [5] in the semantic tasks.

The work of [5] take two sets of vectors, in their case **Word2Vec** vectors, and an known relation between this two sets. They used an lexical dictionary to get this relation. For the missing relations they use a cluster algorithm. After that, they project the vectors using the canonical correlation analysis[2]. The CCA is basically a black box, which gets the best projection between two known vectors.

We propose to use the canonical correlation analyses, same as [7], to help to prediction process, to get right parsing decisions, but instead of combine two "different" languages, we propose to combine sisters languages. Because, hypothetically these vectors would be better then just combine non sister languages.

3 Our architecture

Our architecture is based on [2]. Theirs architecture is simple multi-layer perceptron, with one embeddings layer, to deal with embeddings matrix of the features and one hidden layer to extract the higher level feature representation, and off course one soft-max layer as usual in deep learning.

But, the problem with their architecture is the soft-max layer. They use only three different classes, at the classification level. The architecture only classify in right/left arc and shift operations, without the label. They only get the label in the oracle processes. Our architecture really learn how to classify the $N \times 2 + 1$ labels. We transform these classes in one-hot representation. For example, if we use the universal dependency corpus [13] as gold corpus, we can have 81 different vectors in one-hot form.

An one-hot representation is a binary vector for each word. The idea is the neural network can learn better features representations for each class individually, making the task more complete and accurate then just look the best decision at the oracle.

Our neural dependency parser is based on arc-stadart operations, and we use the same feature template as [17]. This features are a concatenation of words, pos and labels. In our neural architecture we extract the embeddings for each feature, in distributional representation way, similar as [2] and [7]. So, the entrance of our architecture are the concatenation of 46 different tokens/features through the embedding layer. The training objective is maximize the cross entropy plus l_2 term.

But, we not only wants to use this neural network based parser to parse monolingual sentences. Instead, we want to use that architecture to apply language

[2] CCA code: http://www.mathworks.com/help/stats/canoncorr.html

transfer techniques based on distributional correlation vectors, to do multilingual parsing based on sister languages hypothesis.

4 Some thoughts

Today the use of the context information is very important in NLP area, this concept is important to define the meaning of an word, for example. As we shown [5] and [11] demonstrate the importance of learn multilingual distributions or use distributional representations from other languages, to achieve best representations. But, none of that works talk about the importance of obtain this representations based on sister languages. We re-analyze the work of [5]. In his work, [5][3] get very goods results combining german with english, which make sense in the hypothesis of sister languages, and also get results in other languages like french/english and spanish/english. They tested the combined distributions in several tasks like semantic and, even syntactic.

Another interesting work who get very good results using the sister languages hypothesis was [7]. In his work [7] use a neural dependency parser for multilingual parsing using transfer language techniques, as results he shown the improvement of project word vectors from english over german, and then apply in dependency parser task. He gets as result of that technique 60.35 unlabeled score, instead of 0.61 comparing with state of art multilingual parsing [9].

Based on these two analysis we argue in favor of our hypothesis, and his use in parsing in multilingual parsing.

5 Conclusions

We review the literature and shown the importance of learn distributional representations based on sisters languages. We hope that type of construction of representations it‘ll be confederated, to make better representations. But, our efforts are to use that type of information to make betters parsing decisions. To fill that proposition, we proposed an neural architecture based parser. We hope get better or similar results to [2] for monolingual parsing, and [7] for multilingual. We hope that this architecture fill the gap of the statical oracle.

As future work, we can construct an better model for extract contextual information, helping the model to make better parsing decisions. Like, the very interesting work of [4] that uses a recurrent neural technique to extract better transition parsing states. But, their architecture have a very interesting approach of staking LSTM layers to do dependency parsing operations, like push (SHIFT) and pop (right/left arc). And, modeling this states in the neural network can help the parser do better decisions for unsupervised parsing, without the use of static features as input of the network.

[3] Gets an accuracy of of $69,66$ by combining german-english and, $69,61$ for french-english, and $67,76$ for spanish/english.

References

1. Ang Lu, Weiran Wang, M.B.K.G., Livescu, K.: Deep multilingual correlation for improved word embeddings. In: NAACL (2015)
2. Chen, D., Manning, C.D.: A fast and accurate dependency parser using neural networks. In: emnlp (2014)
3. Collobert, R., Weston, J.: A unified architecture for natural language processing: Deep neural networks with multitask learning. In: ICML (2008)
4. Dyer, C., Ballesteros, M., Ling, W., Matthews, A., Smith, N.A.: Transition-based dependency parsing with stack long short-term memory. In: acl (2015)
5. Faruqui, M., Dyer, C.: Improving vector space word representations using multilingual correlation. In: EACL (2014)
6. andJohan Hall, J.N., Nilsson, J.: Maltparser: A data-driven parser-generator for dependency parsing. In: LREC (2006)
7. Jiang Guo, Wanxiang Che, D.Y.H.W.T.L.: Cross-lingual dependency parsing based on distributed representations. In: ACL (2015)
8. Koo, T., Carreras, X., Collins, M.: Simple semi-supervised dependency parsing. In: acl (2008)
9. McDonald, R., Petrov, S., Hall, K.: Multi-source transfer of delexicalized dependency parsers. In: NACL (2011)
10. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. In: ICLR (2013)
11. Minh-Thang Luong, H.P., Manning, C.D.: Bilingual word representations with monolingual quality in mind. In: naacl (2015)
12. Nivre, J.: Universal dependency parsing (2014), http://stp.lingfil.uu.se/~nivre/docs/NivreSPMRL.pdf, nvited talk at SPMRL-SANCL, Dublin.
13. Ryan McDonald, Joakim Nivre, Y.Q.B.Y.G.D.D.K.G.K.H.S.P.H.Z.O.T.C.B.N.B.C.J.L.: Universal dependency annotation for multilingual parsing. In: ACL (2013)
14. Tackstrom, O., McDonald, R., Uszkoreit, J.: Cross-lingual word clusters for direct transfer of linguistic structure. In: NAACL (2012)
15. Tsarfaty, R., Seddah, D., Kubler, S., Nivre, J.: Parsing morphologically rich languages: Introduction to the special issue. In: ACL (2013)
16. Turian, J., Ratinov, L., Bengio, Y.: Word representations: A simple and general method for semi-supervised learning. In: ACL (2010)
17. Zhang, Y., Nivre, J.: Transition-based dependency parsing with rich non-local features. In: ACL (2011)