Neural Language Models vs Wordnet-based Semantically Enriched Representation in CST Relation Recognition

Arkadiusz Janz, Maciej Piasecki, Piotr Wątorski
Wrocław University of Science and Technology, Poland
{arkadiusz.janz|maciej.piasecki|piotr.watorski}@pwr.edu.pl

Abstract

Neural language models, including transformer-based models, that are pre-trained on very large corpora became a common way to represent text in various tasks, including recognition of textual semantic relations, e.g. Cross-document Structure Theory. Pre-trained models are usually fine tuned to downstream tasks and the obtained vectors are used as an input for deep neural classifiers. No linguistic knowledge obtained from resources and tools is utilised. In this paper we compare such universal approaches with a combination of rich graph-based linguistically motivated sentence representation and a typical neural network classifier applied to a task of recognition of CST relation in Polish. The representation describes selected levels of the sentence structure including description of lexical meanings on the basis of the wordnet (plWordNet) synsets and connected SUMO concepts. The obtained results show that in the case of difficult relations and medium size training corpus semantically enriched text representation leads to significantly better results.

1 Introduction

Recognition of semantic relations linking text fragments may provide insight into the semantic-pragmatic structure of text or be a basis for human-like reasoning. The Cross-document Structure Theory (CST) (Radev, 2000) defines a system of semantic relations connecting topically related texts. However, due to the large number of relations and often subtle differences between them, CST relation recognition is known to be much harder than Textual Entailment (TE) recognition. TE depends on a binary decision whether one piece of text semantically entails another one due to their content, while CST is a model of more general use, but more difficult to achieve good results, especially when a classifier is trained on a domain different than the domain of its application.

CST relations are based on relations between semantic content of the text fragments, like *Subsumption* or *Background*. Such semantic oppositions are not trivial in the case of several relation types. For instance, differences in the definitions of *Description*, *Follow-up* or *Elaboration* indicate some potential difficulties that may arise when we want to recognize certain types of relations. In case of Description, the new additional information is about the current, non-historical nature of an event, e.g. the first sentence describes an object or entity appearing in the second sentence. Elaboration provides some additional details regarding the event, but generally the sentences convey the same core information. Follow-up provides some unrevealed facts about the event but appearing after occurrence of this event, thus it may be some kind of description for related events.

Janz et al. (2018) showed that enriched graph based representation of sentences that combines elements from the levels of words, syntactic structures and also semantic structures results in significant improvement of the recognition performance in comparison to less informed approaches of simpler representation models. The semantic parts of graphs included wordnet synsets, SUMO (Pease, 2011) concepts, proper names and selected semantic relations from noun phrases, where the wordnet and SUMO based graph elements dominates. The recent rapid development of approaches based on word embeddings, neural language models and deep neural classifiers shows that novel end-to-end methods can be very successful when applied to downstream tasks. In our work we want to verify this common claim by comparing approaches util-

ising more complex text representation with those based on versatile neural language models or word embeddings. The goal of this paper is to compare two approaches: a typical 'neural' approach and the elaborated method of Janz et al. (2018), as well as their combination, as the first step into this domain. The approach was presented on the basis of a well built, but medium size corpus. This may be an exemplar of a practical problem: in practice many tasks are sparingly illustrated by annotated data, and, thus, poses challenges to 'neural' methods as they require large resources to fine tune representations based on pre-trained neural models (contextual embeddings) to the problem. For the comparison we used the same annotated corpus and exactly the same representation as in (Janz et al., 2018) and neural language models pretrained on very large corpora, and fine tuned on the same annotated corpus. Our aim is to verify a claim that knowledge-based representation, especially wordnet-based, may be still useful in such cases.

2 Related Work

In Zhang et al. (2003) CST relations were recognised by a supervised approach with boosting on the basis of lexical, syntactic and semantic features extracted from sentence pairs. The evaluation was performed in two steps: binary classification for relationship detection, and multi-class classification for relationship recognition. (Zhang and Radev, 2005), in addition to labelled data, exploited also unlabelled instances that improved the performance. Boosting technique was used in combination with the same set of features to classify the data in CSTBank (Radev et al., 2004). Relation detection was significantly improved to F1-score = 0.8839. However, recognition of the relation type was still unsatisfactory.

(Aleixo and Pardo, 2008) is one of few works that address the problem of CST relations recognition for languages other than English. They utilised CST in search for topically related Portuguese documents. They applied a supervised approach based on similarity measures calculated for sentence pairs from different documents: cosine similarity and a variant of the Jaccard index. Cut-off thresholds for the similarity were studied in combination with the performance of classifiers. Aleixo and Pardo (2008) constructed a CST corpus for Portuguese and used it to conduct their study.

Zahri and Fukumoto (2011) applied the supervised learning to recognise a a subset of CST relations: *Identity*, *Paraphrase*, *Subsumption*, *Elaboration* and *Partial Overlap*. SVM algorithm was used and examples from CSTBank. The features of Aleixo and Pardo (2008) were expanded with: cosine similarity of word vectors, Jaccard Index to measure intersection of common words, longer sentence indicator, and uni-directional word coverage ratio.

Kumar et al. (2012a) followed Zahri and Fukumoto (2011), but restricted the set of relations to four and used only four features: tf-idf based cosine sentence similarity, words coverage ratio, sentence length difference, and longer sentence flag. The performance of SVM in relation recognition was between (F1): 0.54 and 0.91 For the same relations Kumar et al. (2012b) presented results obtained with SVM, a Feed-Forward neural network and CBR (Case-based Reasoning). The features of Zahri and Fukumoto (2011) were extended with the Jaccard based similarity of noun phrases and verb phrases from the compared sentences. The best result was achieved with CBR based on the cosine similarity measure: from 0.722 to 0.966.

Due to the ambiguity in the interpretation of certain CST relationships, Maziero et al. (2014) proposed several refinements to CST in order to reduce the ambiguity. They improved definitions by introducing several additional constraints on the co-occurrence of different relations in texts. The CST taxonomy was amended by adding a division based on the form and information content of relations. The improved model was used in evaluation of supervised CST relation recognition The applied features included: sentence length difference, ratio of shared words, sentence position in text, differences in word numbers across PoSs, and the number of shared synonyms between sentences. The J48-based classifier achieved the best average score of 0.403.

In similar task of implicit discourse relation recognition (Cianflone and Kosseim, 2018) used encoder- decoder (RNN) trained directly on character-level data from a large training corpus of annotated relations (reported F1 between 0.3 and 0.8, depending on the relation type). (Bai and Zhao, 2018) used ELMo (Gardner et al., 2017) and subword-level encoding as an input to a stack of a convolutional encoder, and a recurrent encoder and a multiple layer perceptron with softmax layer

as the classifier – F1 between 0.36 and 0.51 was obtained. (Guo et al., 2018) represented input data by pre-trained word embeddings and next trained a neural tensor network on a large corpus of annotated sentences obtaining F1: 0.38 – 0.72.

However, (Ponti and Korhonen, 2017) used topic model word vectors as representation, but also enriched it with features extracted by dependency parser to recognise causal relations between events – a similar task to ours, but narrower.

3 Dataset

For comparison, we utilised exactly the same dataset as in (Kędzia et al., 2017; Janz et al., 2018), i.e. of sentence pairs annotated with CST relations from the KPWr Corpus (Broda et al., 2012), henceforth *WUT CST*. The underlying corpus used to build the dataset contained $11\,949$ complete documents that were clustered and split into groups of 3 news, each including the most similar and potentially related documents. A set of bundles for manual annotation process was prepared – every one with 10 triples $\{D_1, D_2, D_3\}$ of most similar documents, that were randomly assigned to the annotators. Finally, 96 bundles covering more than 2800 documents were analysed in order to discover new instances of CST relations. The imposed similarity structure facilitated searching for sentence pairs linked by a CST relation. Manually annotated pairs of sentences (by at least by 3 annotators each) representing new instances of CST relations formed the gold reference subcorpus introduced for the first time by Kędzia et al. (2017). Each annotator was exploring the corpus independently. The annotators followed the guidelines used for the construction of CSTBank (Radev et al., 2004) adapted to Polish.

However, for the final corpus *WUT CST*[1] we have rejected uncertain CST instances with inconsistent annotations. This means that our *WUT CST* corpus contains only CST instances with almost homogenous annotations assigned by at least $n-1, n>2$ annotators. The final distribution of collected CST instances in our *WUT CST* corpus is presented in Figure 2.

A corpus, with similar distribution of discourse relations linking multiple documents (texts from journals in Brazilian Portuguese), was also introduced in (Cardoso et al., 2011).

We updated the original dataset to eliminate data redundancy and improve its quality by removing noisy sentence pairs. To deal with highly imbalanced class distribution we decided to completely remove specific minor classes as their sample size was too small to prepare a robust and effective supervised model in a supervised setting. The updated dataset is available at `https://clarin-pl.eu/dspace/handle/11321/781`.

[1] `https://clarin-pl.eu/dspace/handle/11321/305`

4 Neural Representation

Successful applications of transformer-based language models in many NLP tasks seem to be grounded in transfer learning methods and intensive model pre-training on large textual corpora. As pre-trained neural language models became very successful pushing the limits in many different natural language tasks we decided to start off with the most popular transformer-based language models and prepare baseline solutions for CST task. To prepare our baseline solutions we decided to choose ELMo and BERT (Devlin et al., 2019) pre-trained language models as it has been shown that they express good performance in Natural Language Inference (NLI) tasks e.g. Textual Entailment (TE). This choice was motivated by the fact that NLI tasks and CST theory are strongly interconnected.

4.1 Pre-trained Language Models

In recent years the general interest in neural language modeling has led to emergence of new pre-trained language models for many different natural languages and Polish language is no exception here. In this paper we used the largest freely available language models pre-trained on selected Polish corpora.

4.2 Multilingual BERT

BERT is a popular and very successful transformer-based architecture for language modeling. It uses masked language modeling with next sentence prediction as an auxiliary objective for training. In this work we use the Multilingual Cased model. The authors used Wikipedia dump extracted for over 100 languages to prepare the model. Still, the language modeling abilities of this model can vary across different languages due to the differences of Wikipedia dump size and thematic representativeness for different languages.

Historical Background
Phoenix wylądował 25 maja 2008 na północnym biegunie Marsa z 3 miesięczną misją badania planety.
Phoenix landed on the Mars' North Pole on 25th May 2008 in a three month mission to explore the planet.

Z tego powodu NASA podejmuje okresowe nasłuchy lądownika.
Due to this reason NASA undertakes periodical listening for the landing module.

Fulfilment
21 lutego 2008 po północy (wg czasu polskiego) miało miejsce całkowite zaćmienie Księżyca.
21st February 2008 past midnight (Polish time) a total Lunar eclipse took place.

21 lutego 2008 po północy (w Polsce) będzie można zaobserwować całkowite zaćmienie Księżyca.
21st February 2008 past midnight (in Poland) it will be possible to observe a total Lunar eclipse.

Follow up
Były premier Leszek Miller będzie kandydował do wyborów parlamentarnych z listy Samoobrony.
The former prime minister Leszek Miller will candidate in parliamentary election from the Samoobrona list.

2007-09-15: Leszek Miller odszedł z SLD
2007-09-15: Leszek Miller left SLD.

Figure 1: Examples of sentence pairs linked by CST relations in WUT CST dataset.

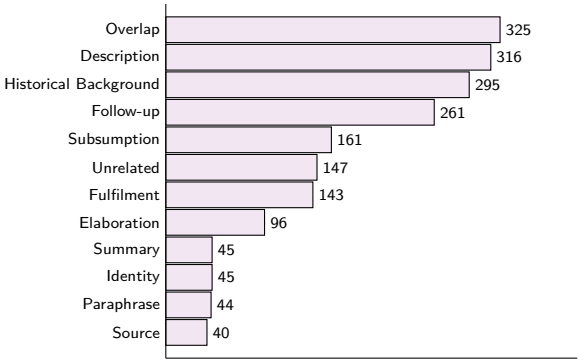

Figure 2: Relations distribution in *WUT CST*.

4.3 Polish BERT-based Models

As it was stated in previous section, the quality of pre-trained language models is mainly dependent on the quality of training corpora. A large part of Polish NLP community in the last few years was focused on adopting well-known language models and training them on publicly available Polish corpora due to the insufficient performance of models trained on Polish Wikipedia only.

HerBERT[2] is a new Polish language model (Rybak et al., 2020) pre-trained on multiple open Polish corpora. The model itself is mainly based on BERT architecture but it also uses dynamic masking as it was originally proposed in RoBERTa (Liu et al., 2019) language model.

4.4 Polish ELMo

ELMo is a language model based on stacked bidirectional LSTM architecture with character-level convolutions. We decided to choose a publicly available model [3] trained on KGR10 corpora as it was the only one model of this kind fully pre-trained on large Polish data from scratch. The KGR10 (Kocoń and Gawor, 2019) is one of the largest Polish corpora of over 4 billion words.

The model was tested extrinsically in selected Polish benchmarks prepared for different NLP tasks e.g. Named Entity Recognition (NER), Sentiment Analysis (SA), or Recognition of Temporal Expressions.

5 Complex Representation

We started from the representations proposed by Janz et al. (2018). In the original work the best result were reported for the combination of manually engineered features and complex graph-based similarities. We preserve the original setting and we shortly recollect features inspired by the literature in Sec. 5.1 and the graph-based features in Sec. 5.2. Finally, both groups of features are concatenated into one single vector as a sentence representation in the experiments discussed in Sec. 8.

[2] https://huggingface.co/allegro/herbert-klej-cased-v1

[3] https://clarin-pl.eu/dspace/handle/11321/690

As graph-based features are those making the differences, cf (Janz et al., 2018), the combined representation will be referred as graph-based representation (or features).

5.1 Bag-of-Elements Representation

The simplest representation of a sentence is a bag of words (a collection of words), i.e. a set of pairs: words plus their frequencies. This basic idea was expanded to bags of diverse elements resulting from rich pre-processing of analysed data.

As it was proposed by Janz et al. (2018) we applied the following pre-processing steps: text lemmatisation and morphosyntactic tagging (Radziszewski, 2013), dependency parsing (Wróblewska and Woliński, 2012; Wróblewska, 2014) in parallel with chunking (Radziszewski and Pawlaczek, 2013), named entity recognition (Marcińczuk et al., 2013), multi-word expression recognition (Radziszewski et al., 2011), and word sense disambiguation (Kędzia et al., 2015; Piasecki et al., 2016). Selected semantic relations inside nominal phrases were recognised by hand-crafted rules (Kedzia and Maziarz, 2013).

The output from word sense disambiguation tool was used to map words to the appropriate synsets of plWordNet 3.0. As plWordNet 3.0 synsets were semi-automatically mapped onto concepts from SUMO ontology (Pease, 2011), thus, the words could be also mapped to their corresponding concepts. We used the metadata obtained by applying aforementioned pre-processing steps to reproduce graph-based representations of the sentences from WUT CST dataset.

5.2 Graph-based Representation

The graph-based representation proposed by Janz et al. (2018) represents a single sentence as a collection of graphs where the nodes correspond to the elements of the detected linguistic structure (e.g. words, lemmas, senses, or ontology concepts) and links reflect relations held between these elements. Concerning the latter a relation can be a simple linear precedence in text, but also a syntactic or semantic link recognised by an appropriate tool. All graphs used are directed.

Some of them include elements external to the sentence structure originating from the linked knowledge resources, e.g. an ontology. A pair of sentences may be described not only by a pair of graphs themselves, but also by values of different similarity measures defined on their graphs.

Many graph types were generated and used in (Janz et al., 2018) by combining different types of nodes with a variety of edge types. Four node types are used:

1. *Lemma* – a graph node represents a lemma of the word w_i converted to lowercase; all words from the sentence with the same lemma (irrespectively of PoS) are represented by the same node;

2. *Lemma PoS* – a node represents a lowercased lemmas, but concatenated with the PoS label, e.g. the Polish word *piec* can be morphologically disambiguated as a verb or noun *Kasia piecze:v ciasto w piecu:n* 'Kasia is baking a cake in the oven'. Using *Lemma lower* type, the words *piecze* '[he/she] bakes' and *piecu* 'an oven:inst' will be represented by a single node labelled as *piec*, while in *Lemma PoS lower* type there will be two different nodes: *piec:n* and *piec.v*.

3. *Synset* – a node represents a plWordNet synset of a given word; the synsets are obtained by applying word sense disambiguation tool to input sentences,

4. *Concept* – a node is a SUMO concept identified on the basis of the disambiguated synset of a word and its mapping to a SUMO concept (Kędzia and Piasecki, 2014).

The edge types originate from the automatically recognised lexical and semantic relations in a sentence. The edge direction reflects the original link direction:

word order – edges represent the word order,

head order – an edge represents the relative order of the heads of *agreement phrases* in a sentence – phrases and their heads are recognised by IOBBER chunker, edges signal the linear order of the heads,

NE order – similar to the head and word orders, but it represents the linear order of the named entities *NE* in a sentence,

syntactic dependency – represents the dependency relations, recognised by the Polish Malt parser (Wróblewska and Woliński, 2012),

nominal structure relations – similar to the *syntactic dependency*, but relations come from *Defender* parser based on IOBBER and introduce deeper syntactic-semantic relation structures into the representation of NPs, cf (Kedzia and Maziarz, 2013).

semantic role – represents semantic roles from *NPSemrel*[4], a Polish shallow semantic parser (Kedzia and Maziarz, 2013), e.g. *agent*, *theme*.

An example sentence with one of its graph representations is presented in Fig. 3. The lemmas were replaced with the equivalent *Synset* nodes from plWordNet after disambiguating them with word sense disambiguation tool.

Constructed graphs can be enriched and generalised to some extent by expanding them with additional nodes from the linked semantic resources. Janz et al. (2018) used for this purpose plWordNet 3.0 and SUMO ontology. For all node pairs from the original graph the shortest paths going across given semantic resource are identified and then included into the expanded graph, cf (Janz et al., 2018). This means that the additional nodes are included together with the resource-specific relations comprising the paths.

All types of edges and nodes and their combinations characterised above were used for the description of pairs of sentences in the experiments by (Janz et al., 2018) and also in ours[5]. As a result, 12 possible graph types in total can be generated, i.e. 4 types of nodes and 3 types of resource expansion, namely: *Lemma lower* graph expanded with SUMO, *Lemma PoS lower* expanded on the basis of plWordNet, *Concept* expanded with SUMO (additional structures, generalisation by higher level concepts) and *Synset* graph expanded with both plWordNet and SUMO (as one connected semantic network). To generate the features we used all of possible graphs we could obtain with this procedure.

Graphs created for a pair of sentences – a training/testing case – were mainly used to calculate their similarity. The computed values of similarity measures were included in vector space representation describing given classification case. To compute similarities six different measures were applied Janz et al. (2018):

1. *Graph Edit Distance* (Fernández and Valiente, 2001) (GED) – the minimal sum of the costs of atomic operations transforming one graph into the other;

2. *Maximum Common Subgraph* (MCS) (Bunke and Shearer, 1998) – the size of maximum common subgraph normalised by the size of the bigger graph;

3. Measure *WGU* (Wallis et al., 2001) – the size of the maximum common subgraph normalised by the sum of the sizes of both graphs minus it.

4. *UGU* (Bunke, 1997) is simply $|G_1| + |G_2| - 2 \cdot |mcs(G_1, G_2)|$, where G_1 and G_2 are sentence graphs, and $mcs(...)$ returns the maximum common subgraph.

5. *MMCS* Fernández and Valiente (2001) expresses the dissimilarity of graphs G_1 and G_2: $|MCS(G_1, G_2)| - |mcs(G_1, G_2)|$.

6. *Contextual BOW* – based on the application of the *Jaccard* measure to sets of nodes of both graphs expanded with their direct neighbour nodes (Janz et al., 2018).

The calculated similarity values are next used as features – elements of input vectors – during training a classifier. By changing the way of constructing the graphs and computing their similarity we are able to control the representation of sentences in classification process and put more attention to characteristic properties of textual semantic relations (CST relations). This could be a possible way to tune the models by pre-selecting graph representations for the downstream task. However, in this work we do not attempt to perform any tuning procedure using prior graph selection.

[4]The construction of *NPSemrel* is based on hand-written lexicalised syntactic-semantic constraints. They mostly express high precision, i.e. around 60% in the worst cases, but the majority of them is close to 100%. However, the recall is much lower, so F1 measure is typically around 0.5, see (Kedzia and Maziarz, 2013).

[5]More specifically, for every single sentence pair we combine all of possible graph configurations (including possible expansions i.e. plWordNet and SUMO) with all available similarity metrics that can be used to generate similarity-based features. The possible graph configurations were generated in a following way: {[*Lemma*], [*Lemma PoS*], [*Synsets*], [*Concepts*], [*Lemma – plWordNet exp.*], [*Lemma – SUMO exp.*], [*Lemma – plWordNet & SUMO exp.*], [*Lemma PoS – plWordNet exp.*], [*Lemma PoS – SUMO exp.*], [*Lemma PoS – plWordNet & SUMO exp.*], [*Synsets – plWordNet exp.*], ..., [*Concepts – plWordNet & SUMO exp.*]} and so on.

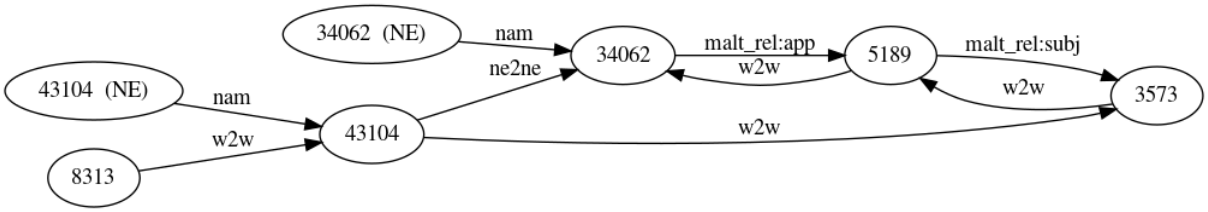

Figure 3: Graph built for the example sentence with *Synset* node type and full set of edges types (*w2w* – word order, *ne2ne* – NE order) (Janz et al., 2018).

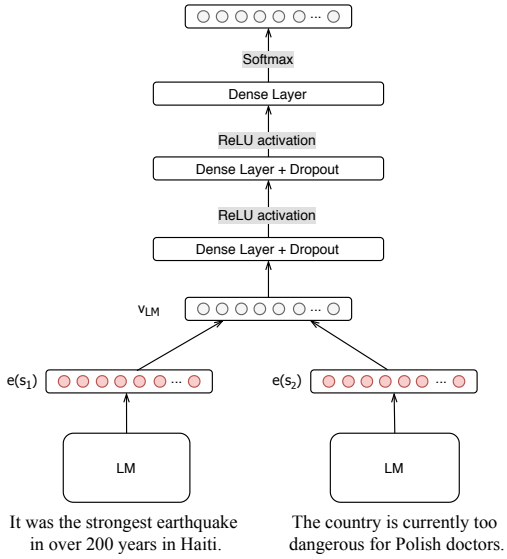

Figure 4: Baseline architecture with transformer-based language modeling and feed forward neural network with multiple dense layers. The language model is used to generate sentence embeddings.

6 CST Relation Recognition

In this section we describe the architecture of the baseline solutions as well as the architecture of their extensions. The architectures are generally based on contextual word embeddings computed by applying pre-trained language models to given sentence pairs. Our main aim was to evaluate existing modern language models and compare them with various wordnet-based features in the task of the recognition of discourse relations. As the task itself is closely related to other NLI tasks, we had assumed that the applied neural language models should bring very good results.

The first architecture uses contextual word embeddings to generate sentence embeddings of sentence pairs from the WUT CST corpus. Given a sentence pair (s_1, s_2) we generate an input vector space representation of this pair $v_{LM} \in \mathcal{R}^{2d_{LM}}$ by concatenating the representations of its sentences $e(s_1) \in \mathcal{R}^{d_{LM}}$, $e(s_2) \in \mathcal{R}^{d_{LM}}$ computed by a given language model LM. The concatenated vector $v_{LM} = [e(s_1), e(s_2)]$ is then passed through a multi-layer dense classification network with Dropout and ReLU activations on its hidden layers, and Softmax in the output layer. The baseline architecture is presented in figure 4.

Since the architecture itself is very simple we are easily able to extend it and incorporate supplementary features by concatenating precomputed vector space representations of input sentences v_{LM} with a vector v_{GF} of additional features coming from the graph-based representation (including similarity values calculated for various graphs) $v_{input} = [v_{LM}, v_{GF}]$. As a result the baseline architecture is expanded with pre-computed graph-based features mentioned in section 5.2.

7 Experimental Setting

To conduct the experiments we used the updated version of *WUT CST* dataset as it was mentioned in Sec. 3. We divided the dataset into three distinct parts to train, tune and evaluate selected neural models and their extensions. To prepare and test the models we applied popular transformer library called Hugging Face[6] Most of the language models used in this work were fine-tuned to the task to obtain the best possible results. We found that fine-tuning the models slightly increases their performance. The ELMo appeared to be difficult to tune, thus, we decided to test only the pre-trained version of this model (ELMo$_{nFT}$). For each pair of sentences we compute their vectors using given language model and classify them with the same baseline architecture presented in figure 4. The extended models used additional graph-based features as an input to classification network (see Sec. 6). As a baseline approach we used Logistic Model Tree (LMT) (Landwehr et al., 2005) trained on graph-based features only as it was proposed

[6] https://huggingface.co

Model	*Overlap*	*Description*	*Background*	*Follow-up*	*Subsumption*	*Unrelated*	*Fulfilment*	*Elaboration*	*Summary*	*Identity*	*Paraphrase*	*Source*	*Accuracy*
BERT	0.35	**0.89**	0.78	0.56	0.40	0.61	0.62	0.25	0.13	0.84	0.15	0.31	0.58
RoBERTa	0.41	0.85	0.83	0.61	0.55	0.68	0.57	0.32	0.21	1.00	0.47	0.36	0.63
HerBERT	0.37	0.84	0.78	0.55	0.30	0.62	0.52	0.14	**0.22**	0.63	0.00	0.29	0.57
ELMo$_{nFT}$	0.36	0.76	0.72	0.51	0.28	0.69	0.67	0.26	0.00	0.50	0.00	0.00	0.55
$GF-$LMT	0.69	0.68	0.83	0.74	0.75	0.95	0.60	0.40	0.00	0.95	0.53	**0.75**	0.71
$GF-$BERT	**0.82**	0.77	**0.86**	**0.88**	0.76	**0.97**	0.54	0.00	0.00	**1.00**	0.53	0.67	**0.78**
$GF-$RoBERTa	0.82	0.76	0.76	0.87	**0.78**	0.95	0.68	0.17	0.00	0.89	**0.55**	0.67	0.74
$GF-$HerBERT	0.79	0.85	0.81	0.86	0.71	0.88	0.79	0.15	0.00	0.84	0.36	0.67	0.77
$GF-$ELMo$_{nFT}$	0.80	0.80	0.84	0.87	0.65	0.87	**0.76**	**0.42**	0.20	0.86	0.36	0.67	0.77

Table 1: F1-scores of evaluated solutions computed with respect to CST relation type. The last column presents the final accuracy of the models.

in (Janz et al., 2018). We selected the default parameters offered by WEKA framework (Hall et al., 2009).

8 Results

The overall results are presented in Table 1 which includes the final F1-scores of four baseline language models, as well as their versions expanded with graph-based representation – marked by GF prefix. They are compared to graph-based only baseline solution $GF-$LMT – using Logistic Model Trees as a classifier and graph-based representation only. The baseline $GF-$LMT model, identical to the one of (Janz et al., 2018) achieved significantly better results, especially for many under-represented classes. The language models were fine-tuned multiple times to our task to ensure that we obtain the best possible results. The language models enhanced with the same graph-based features as our baseline model – $GF-$BERT, $GF-$HerBERT, $GF-$RoBERTa, and $GF-$ELMo appeared to beat their initial results as it was expected.

9 Conclusions

Neural language models (word and sentence embeddings) are capable to express enormous amounts of knowledge about possible language contexts, if pre-trained on a corpus that is large enough and representative. We applied models which have been built on very large corpora and showed very good performance when used as a basis in many applications. However, the complexity of such pre-trained models causes that machine learning algorithm must cope with it, unless they are fine tuned to a given problem on a dataset large enough. In order to do this, one requires appropriate data, both in terms of the good representation of the problem, and, very important, substantial size. Extraction of elements of linguistic structures introduces generalisation, highlighting most important markers and a kind of mapping to an abstract space. We showed that such enriched representation may help in problems where we do not have enough training data. A future challenge is to find a way of balancing and combining the two approaches.

Acknowledgments

The work financed as part of the investment in the CLARIN-PL[7] research infrastructure funded by the Polish Ministry of Science and Higher Education.

References

Priscila Aleixo and Thiago Alexandre Salgueiro Pardo. 2008. Finding Related Sentences in Multiple Documents for Multidocument Discourse Parsing of Brazilian Portuguese Texts. In *Companion Proceedings of the XIV Brazilian Symposium on Multimedia and the Web*, WebMedia '08, pages 298–303. ACM, New York, NY, USA.

Hongxiao Bai and Hai Zhao. 2018. Deep enhanced representation for implicit discourse relation recognition. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 571–583. Association for Computational Linguistics, Santa Fe, New Mex-

[7] `http://clarin-pl.eu`

ico, USA. URL https://www.aclweb.org/anthology/C18-1048.

Bartosz Broda, Michał Marcińczuk, Marek Maziarz, Adam Radziszewski, and Adam Wardyński. 2012. KPWr: Towards a Free Corpus of Polish. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. European Language Resources Association (ELRA), Istanbul, Turkey.

H. Bunke. 1997. On a Relation Between Graph Edit Distance and Maximum Common Subgraph. *Pattern Recogn. Lett.*, 18(9):689–694.

Horst Bunke and Kim Shearer. 1998. A Graph Distance Metric Based on the Maximal Common Subgraph. *Pattern Recogn. Lett.*, 19(3-4):255–259.

Paula C.F. Cardoso, Erick G. Maziero, María Lucía Castro Jorge, Eloize R.M. Seno, Ariani Di Felippo, Lucia Helena Machado Rino, Maria das Gracas Volpe Nunes, and Thiago Alexandre Salgueiro Pardo. 2011. CSTNews - A discourse-annotated corpus for single and multi-document summarization of news texts in Brazilian Portuguese. In *Proceedings of the 3rd RST Brazilian Meeting*, pages 88–105. Cuiabá, Brazil.

Andre Cianflone and Leila Kosseim. 2018. Attention for implicit discourse relation recognition. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA), Miyazaki, Japan. URL https://www.aclweb.org/anthology/L18-1306.

J. Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

Mirtha-Lina Fernández and Gabriel Valiente. 2001. A Graph Distance Metric Combining Maximum Common Subgraph and Minimum Common Supergraph. *Pattern Recogn. Lett.*, 22(6-7):753–758.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke S. Zettlemoyer. 2017. Allennlp: A deep semantic natural language processing platform.

Fengyu Guo, Ruifang He, Di Jin, Jianwu Dang, Longbiao Wang, and Xiangang Li. 2018. Implicit discourse relation recognition using neural tensor network with interactive attention and sparse learning. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 547–558. Association for Computational Linguistics, Santa Fe, New Mexico, USA. URL https://www.aclweb.org/anthology/C18-1046.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. 2009. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18.

Arkadiusz Janz, Paweł Kędzia, and Maciej Piasecki. 2018. Graph-based complex representation in inter-sentence relation recognition in polish texts. *Cybernetics and Information Technologies Journal*, 18(1):152–170.

Pawel Kedzia and Marek Maziarz. 2013. Recognizing semantic relations within Polish noun phrase: A rule-based approach. In *RANLP*.

Paweł Kędzia, Maciej Piasecki, and Arkadiusz Janz. 2017. Graph-based approach to recognizing CST relations in Polish texts. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 363–371. INCOMA Ltd., Varna, Bulgaria. URL https://doi.org/10.26615/978-954-452-049-6_048.

Jan Kocoń and Michał Gawor. 2019. Evaluating KGR10 polish word embeddings in the recognition of temporal expressions using BiLSTM-CRF. *arXiv preprint arXiv:1904.04055*.

Yogan Jaya Kumar, Naomie Salim, Ahmed Hamza, and Albarraa Abuobieda. 2012a. *Automatic identification of cross-document structural relationships*, pages 26–29.

Yogan Jaya Kumar, Naomie Salim, and Basit Raza. 2012b. Cross-document Structural Relationship Identification Using Supervised Machine Learning. *Appl. Soft Comput.*, 12(10):3124–3131.

Paweł Kędzia and Maciej Piasecki. 2014. Ruled-based, Interlingual Motivated Mapping of plWordNet onto SUMO Ontology. In Nicoletta

Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asunción Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014), Reykjavik, Iceland, May 26-31, 2014.*, pages 4351–4358.

Paweł Kędzia, Maciej Piasecki, and Marlena Orlińska. 2015. Word sense disambiguation based on large scale Polish CLARIN heterogeneous lexical resources. *Cognitive Studies / Études cognitives*, 15:269–292. URL `https://ispan.waw.pl/journals/index.php/cs-ec/article/download/cs.2015.019/1765`.

Niels Landwehr, Mark Hall, and Eibe Frank. 2005. Logistic model trees. *Machine learning*, 59(1-2):161–205.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Michał Marcińczuk, Jan Kocoń, and Maciej Janicki. 2013. Liner2 – a customizable framework for proper names recognition for Polish. In Robert Bembenik, Łukasz Skonieczny, Henryk Rybiński, Marzena Kryszkiewicz, and Marek Niezgódka, editors, *Intelligent Tools for Building a Scientific Information Platform*, pages 231–253.

Erick Galani Maziero, Maria Lucía Del Rosário Castro Jorge, and Thiago Alexandre Salgueiro Pardo. 2014. Revisiting Cross-document Structure Theory for Multi-document Discourse Parsing. *Inf. Process. Manage.*, 50(2):297–314.

Adam Pease. 2011. *Ontology: A Practical Guide*. Articulate Software Press, Angwin, CA.

Maciej Piasecki, Pawel Kędzia, and Marlena Orlińska. 2016. plWordNet in Word Sense Disambiguation task. In *GWC 2016, Proceedings of the 8th Global Wordnet Conference, Bucharest, 27-30 January 2016 Osaka, Japan*, pages 280–290.

E. Ponti and A. Korhonen. 2017. Event-related features in feedforward neural networks contribute to identifying implicit causal relations in discourse. In *Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics*, pages 25–30.

Dragomir R. Radev. 2000. A Common Theory of Information Fusion from Multiple Text Sources Step One: Cross-document Structure. In *Proceedings of the 1st SIGdial Workshop on Discourse and Dialogue - Volume 10*, SIGDIAL '00, pages 74–83. Association for Computational Linguistics, Stroudsburg, PA, USA.

Dragomir R. Radev, Jahna Otterbacher, and Zhu Zhang. 2004. Cst bank: A corpus for the study of cross-document structural relationships. In *LREC*. European Language Resources Association.

Adam Radziszewski. 2013. *A Tiered CRF Tagger for Polish*, pages 215–230. Springer Berlin Heidelberg, Berlin, Heidelberg. URL `https://doi.org/10.1007/978-3-642-35647-6_16`.

Adam Radziszewski and Adam Pawlaczek. 2013. *Language Processing and Intelligent Information Systems: 20th International Conference, IIS 2013, Warsaw, Poland, June 17-18, 2013. Proceedings*, chapter Incorporating Head Recognition into a CRF Chunker, pages 22–27. Springer Berlin Heidelberg, Berlin, Heidelberg.

Adam Radziszewski, Adam Wardyński, and Tomasz Śniatowski. 2011. WCCL: A morpho-syntactic feature toolkit. In *Proceedings of the Balto-Slavonic Natural Language Processing Workshop (BSNLP 2011)*. Springer.

Piotr Rybak, Robert Mroczkowski, Janusz Tracz, and Ireneusz Gawlik. 2020. Klej: Comprehensive benchmark for polish language understanding. *arXiv preprint arXiv:2005.00630*.

W. D. Wallis, P. Shoubridge, M. Kraetz, and D. Ray. 2001. Graph Distances Using Graph Union. *Pattern Recogn. Lett.*, 22(6-7):701–704.

Alina Wróblewska and Marcin Woliński. 2012. *Preliminary Experiments in Polish Dependency Parsing*, pages 279–292. Springer Berlin Heidelberg, Berlin, Heidelberg.

Alina Wróblewska. 2014. *Polish Dependency Parser Trained on an Automatically Induced Dependency Bank*. Ph.D. dissertation, Institute of Computer Science, Polish Academy of Sciences, Warsaw.

Nik Adilah Hanin Binti Zahri and Fumiyo Fukumoto. 2011. *Multi-document Summariza-*

tion Using Link Analysis Based on Rhetorical Relations between Sentences, pages 328–338. Springer Berlin Heidelberg, Berlin, Heidelberg.

Zhu Zhang, Jahna Otterbacher, and Dragomir Radev. 2003. Learning Cross-document Structural Relationships Using Boosting. In *Proceedings of the Twelfth International Conference on Information and Knowledge Management*, CIKM '03, pages 124–130. ACM, New York, NY, USA.

Zhu Zhang and Dragomir Radev. 2005. Combining Labeled and Unlabeled Data for Learning Cross-document Structural Relationships. In *Proceedings of the First International Joint Conference on Natural Language Processing*, pages 32–41. Springer-Verlag, Berlin, Heidelberg.