Graph-Based Approach to Recognizing CST Relations in Polish Texts

Paweł Kędzia
Wrocław University
of Science and Technology

Maciej Piasecki
Wrocław University
of Science and Technology

Arkadiusz Janz
Wrocław University
of Science and Technology

{pawel.kedzia,maciej.piasecki,arkadiusz.janz}@pwr.edu.pl

Abstract

This paper presents a supervised approach to the recognition of Cross-document Structure Theory (CST) relations in Polish texts. In the proposed, graph-based representation is constructed for sentences. Graphs are built on the basis of lexicalised syntactic-semantic relations extracted from text. Similarity between sentences is calculated on their graphs, and the values are used as features to train the classifiers. Several different configurations of graphs, as well as graph similarity methods were analysed for this task. The approach was evaluated on a large open corpus annotated manually with 17 types of selected CST relations. The configuration of experiments was similar to those known from SEMEVAL and we obtained very promising results.

1 Introduction

Among large volumes of data available one can find a lot of redundant information, eg. supplementing, overlapping etc. Manual aggregating and synthesizing valuable information from a massive input is laborious. The aim of multi-document discourse parsing is to discover the relations or dependencies linking text passages. The relation we are aiming for are not limited only to the relations between event descriptions. Recognition of discourse relationships linking texts can be useful in many information retrieval applications, and may help in information management.

The Cross-document Structure Theory (CST) (Radev, 2000) introduces an organized structure of semantic links connecting topically related texts. CST relations recognised correctly for text fragments provide a map of the document(s) semantic structure and, e.g., can support multi-document summarization (Kumar et al., 2014). 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.

Our goal is to build a tool for the recognition of CST relations in Polish texts. Firstly, we limited the problem to recognition of relations between sentence pairs, that is even a harder task because of the limited text material. to be processed. For training we used a part of the KPWr Corpus (Broda et al., 2012) based on Polish Wikinews[1]. In the work presented here, we focus on the 17 relations with the largest coverage in the corpus.

2 Related Works

In (Zhang et al., 2003) CST relations were recognized by a supervised approach with boosting on the basis of simple, 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. This idea was expanded Zhang and Radev (2005) by leveraging both labeled and unlabeled data. The exploitation of unlabeled instances 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 F-score = 0.8839. However, recognition of the relation type was still unsatisfactory.

Aleixo and Pardo (2008) is one of a few works that address recognition CST relations for languages other than English. They utilised CST in search for topically related Portuguese documents. They applied a supervised approach based on sim-

[1] https://pl.wikinews.org

Proceedings of Recent Advances in Natural Language Processing, pages 363–371,
Varna, Bulgaria, Sep 4–6 2017.
https://doi.org/10.26615/978-954-452-049-6_048

ilarity 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.

Zahri and Fukumoto (2011) applied the supervised learning to identify a limited set of CST relations: *Identity*, *Paraphrase*, *Subsumption*, *Elaboration* and *Partial Overlap*. They were used in the multi-document summarization task. SVM algorithm was used and examples from CSTBank. The features of (Aleixo and Pardo, 2008) were expanded with: (i) cosine similarity of word vectors, (ii) intersection of common words measured with the Jaccard Index, (iii) an indicator of longer sentence and (iv) one-sided word coverage ratio.

Kumar et al. (2012a) restricted the set of relations further down to four: *Identity*, *Subsumption*, *Overlap* and *Elaboration*. Four features were used: (i) tf-idf based cosine sentence similarity, (ii) words coverage ratio, (iii) sentence length difference and (iv) the indicator of longer sentence. The best performance of SVM in relation recognition was: for *Identity* $F = 0.91$, *Subsumption* 0.59, *Elaboration* 0.54, and 0.62 for *Overlap*. For the same relations Kumar et al. (2012b) presented results obtained with SVM, a Feed-Forward neural network and CBR. The features of (Zahri and Fukumoto, 2011) were extended with the Jaccard based similarity of noun phrases and verb phrases. CBR based on the cosine similarity measure expressed improved results than in (Kumar et al., 2012a): *Identity* 0.966, *Subsumption* 0.803, *Description* 0.786, and 0.722 for *Overlap*.

(Maziero et al., 2014) proposed several refinements to CST in order to reduce the ambiguity. They improved definitions by several additional constraints on the co-occurrence of different relations in texts. The CST taxonomy was amended by introducing a division based on the form and information content of relations. The improved model was used in evaluation of supervised CST relation recognition in three different settings: binary, multi-class and hierarchical (facilitating the proposed taxonomy of relations). The applied features included: sentence length difference, ratio of shared words, sentence position in text, differences of word numbers across PoSs, and the number of shared synonyms between sentences. SVM, Naive Bayes and J48 decision tree were used for classification with the best score of J48. The average F-measure for multi-class scheme was 0.403, while for the binary scheme: 0.673. (without the final decision) and for the hierarchical: 0.724.

3 Dataset

We utilised a dataset of sentence pairs annotated with CST relations from the KPWr Corpus. The corpus consists of complete documents that were grouped by their similarity into groups of 3 news each. The groups include the most similar, potentially topically related documents. The imposed similarity structure facilitated searching for sentence pairs linked by a CST relation. A corpus, with similar distribution of discourse relations linking multiple documents, was also introduced in (Cardoso et al., 2011). It was built from texts from journals in Brazilian Portuguese.

Selected sentences from our corpus were manually annotated with CST relations at least by 3 annotators (linguists) each. Each annotator was exploring the corpus independently, in order to find and annotate inter-document relations inside document groups linking text fragments. The annotators followed the guidelines of CSTBank (Radev et al., 2004) slightly adapted to Polish.

4 Features in Classification

4.1 Baseline Features

As a starting point we used the set features proposed in (Maziero et al., 2014). Our set includes commonly-used, lexical, syntactic and semantic features that were applied for the detection and recognition of CST relationships in supervised approaches. They focus on the grammatical forms in and properties of the linked sentences:

- Shared lemmas – the number of lemmas shared by two sentences,
- Shared PNs – the number of Proper Names shared by two sentences,
- Longest Common Substring – the length of the longest common continuous sub-string of word forms from the two sentences,
- Longest Common Subsequence – the length of the longest common sub-sequence, but the sequences can be discontinuous (i.e. sequence elements can be separated),
- Cosine similarity – the cosine similarity of vectors of the frequency of lemmas,

- Is Longer – equals 1 if the first sentence is longer, 0 for equal, -1 if the second is longer,
- Shared synsets – the number of synsets shared by the two sentences which is normalized by the number of all synsets in the shorter sentence (to make the feature insensitive to sentence length differences),
- PoS similarity – cosine measure of vectors of the frequencies of different Part of Speech in both sentences (4 basic PoS were used),
- SVO Index – the Jaccard Index calculated for vectors of frequencies of triples: subject, verb, object for both texts.

These features were used as a baseline model for the description of text pairs, and compared later with the graph-based representation proposed in the following subsections. Several language tools were used to enrich texts for feature extraction: *Morfeusz* (Woliński, 2006) – a morphological analysis, *WCRFT* (Radziszewski, 2013) – tagger, *Liner2* (Marcińczuk et al., 2013) – recognition of Proper Names, *Maltparser* (Nivre et al., 2007) adapted to Polish (Wróblewska, 2014), *WCCL* (Radziszewski et al., 2011) – recognition of multi-word expressions from plWordNet (Maziarz et al., 2016; Piasecki et al., 2009), *WoSeDon* (Kędzia et al., 2015; Piasecki et al., 2016) – Word Sense Disambiguation, *IOBBER* (Radziszewski and Pawlaczek, 2013) – a syntactic chunker, *Fextor* (Broda et al., 2013) – tool for feature extraction.

4.2 Graph-based Features

The baseline features do not take into account the linguistic structure of the compared sentences. As the parser for Polish has limited accuracy, instead of depending only on the dependency structure produced by the parser we propose a graph-based representation of a sentence (or text) which is flexible and can accommodate results of processing by different language tools.

4.2.1 Graph-based Sentence Representation

Each sentence S_i is represented as a directed graph G_i. Thus, a relation $R(S_1, S_2)$ between sentences S_1 and S_2 is represented as a relation R between graphs G_1 and G_2: $R(G_1, G_2)$. For them we will calculate a similarity value $v_{sim} = SIM(G_i, G_j)$ where SIM means one of the similarity measures discussed in Sec. 4.2.2. Formally, a directed graph $G = (V, E)$ where V is a set of vertices and E is set of directed and ordered edges e A directed edge $e = (n_s, n_t)$ where n_s is the source node and n_t is the target node, the direction is from n_s to n_t. The graphs are built in three steps: creation of nodes and edges on the basis of a sentence and merging the graph with subgraphs extracted from external knowledge sources, i.e. plWordNet and SUMO Ontology (Pease, 2011).

In **the first step** an example sentence pair (S_i and S_j) for a relation R is converted into two separate null graphs, respectively: G_i and G_j. Their nodes are of a selected type T (the same for both graphs), represent the words from the sentences and are not connected to each other. If we select more than one node type, we would obtain several null graphs for each sentence. Depending on the chosen type T_i of node, one or more words from S_i could be represented by the same node:

- *Lemma lower* – this is the simplest node type, a node $n_i \in G_j$ represents a lemma from S_j, which is converted to lowercase. All words from a sentence with the same lemma (irrespectively of PoS) are represented by the same node, e.g., for *Z ogrodu zoologicznego we Wrocławiu uciekł wąż Boa Dusiciel i przemieszcza się w stronę Ostrowa Tumskiego.*
  we obtain the following null graph:
  `{{w1:z},{w2:ogród},{w3:uciec},`
  `{w4:zoologiczny},{w5:wąż},...}`
- *Lemma PoS lower* – in a similar way to *Lemma lower*, nodes represent lowercased lemmas, but PoS label is concatenated, e.g. *cat:n* or the Polish word *piec* can be morphologically disambiguated as a verb or noun *Kasia piecze:v ciasto w piecu:n*. Using *Lemma lower* type, the words *piecze* and *piecu* 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*. For S_{sample} the node of the type *Lemma PoS lower* are: `{{w1:z-prep},{w2:ogród-subst},`
  `{w3:uciec-praet},{w4:wąż-subst},...}`
- *Synset* – nodes represent plWordNet synsets assigned to the words in a sentence as their lexical meanings by WoSeDon, For S_{sample} and the *Synset* node type, the generated null graph consists of :

`{{w1:ogród-4772},{w2:uciec-3573},`
`{w3:zoologiczny-8748},...}`

- *Concept* – nodes are concepts from SUMO Ontology. The concepts are assigned to words in a sentence on the basis of synsets recognised by WoSeDon and the mapping between plWordNet and SUMO (Kędzia and Piasecki, 2014). The null graph of *Concept* type for S_{Sample} is: `{{w1:subsumed-CultivatedLandArea},{w2:subsumed-Attribute},{w3:subsumed-Reptile}{w4:equivalent-Snake},...}`

In **the second step** the null graph constructed in the first step is expanded by adding edges between nodes. If we have multiple null graphs with different node types, we need to expand every null graph from the first step with new edges. The edge types are derived from automatically recognised lexical and semantic relations in a sentence. The e_{type} direction depends on the kind of the relation represented:

- w2w – edges represent the word order in a sentence (*word to word*). If a word w_1 occurs in a sentence before word w_2, then there is a directed edge from w_1 to w_2: $e_{w2w} : (w_1, w_2)$.
- h2h – *head to head* represents the relative order of the heads of *agreement phrases* in a sentence. Each sentence is divided into chunks of three types: Verb Phrase *VP*, Noun Phrase *NP* and Adjective Phrase *AdjP*, that are next subdivided into smaller, *Agreement Phrases* (*AgP*). The relation $h2h$ represents the order of *AgPs* heads. If a AgP head w_{hi} occurs in a sentence before the AgP head w_{hj} then the edge is directed from w_{hi} to w_{hj}: $e_{h2h} : (w_{hi}, w_{hj})$.
- ne2ne – an edge type similar to *w2w* and *h2h*, but in which edges represent the order of the named entities *NE* in a sentence. If named entity w_{nei} occurs before w_{nej} in sentence S, then a directed edge: $e_{ne2ne} : (w_{nei}, w_{nej})$, is added to the graph.
- malt – edges of this type represent the dependency relations. Each dependency relation between two words w_i and w_j, is modelled in the graph as a directed edge with the same direction. If there is a dependency relation $dep_{rel}(w_i, w_j)$, then it is added into the graph as a directed edge with the same direction dep_{rel}: $e_{dep_{rel}}(w_i, w_j)$.
- defender – the type similar to the *malt*, but relations come from *Defender* parser which is based on IOBBER chunker (Kedzia and Maziarz, 2013). Provides deeper relation structures for NPs. We used *malt* and *defender* relations, because in some situations the relations proposed by Malt are incorrect. If there is a dependency for two words w_i and w_j from *Defender*, then it is added as a directed edge to graph: $e_{def}(w_i, w_j)$.
- semantic roles – edges marked as $srole$ represent semantic roles from *NPSemrel*, a Polish shallow semantic parser (Kedzia and Maziarz, 2013). The dependencies proposed by *Defender* are named with semantic roles e.g. *agent*, *theme*. If semantic role is assigned to a pair of words: w_i and w_j, a directed edge is added between the nodes representing w_i and w_j: $e_{srole} : (w_i, w_j)$. The edge is labeled with the semantic role.

All types of edges and nodes were used in our experiments. A single graph G_i represents sentence S_i and contains the edges $E_i \in \{w2w, h2h, ne2ne, malt, def, srole\}$. A graph for sentence $S_{example}$, with *Concept* nodes and full set of possible edge types is shown in Fig. 1.

In **the third step** the constructed graphs are merged with a subgraph extracted from an *External Knowledge Graph* (henceforth *EKG*). Our idea is to add to the graphs built from sentences, more semantic information, extracted from *EKG*. Let G will be a graph with node type t built for sentence S during *second step*, $G = (V_t, E \in \{w2w, h2h, ne2ne, malt, def, srole\})$. EKG_{plwn} is a graph built from plWordNet, where the nodes in $EKG(plwn)$ are the synsets from plWordNet, the edges in $EKG(plwn)$ are the relations from plWordNet. $EKG_{S(plwn)}$ is a subgraph of EKG_{plwn}. EKG_{sumo} is the graph built from SUMO Ontology, where nodes represent concepts from SUMO. The edges in EKG_{sumo} correspond to SUMO relations, and $EKG_{S(sumo)}$ is a subgraph of EKG_{sumo}. A subgraph of EKG is extracted from the source in the following way: for each word w in sentence S we identify the corresponding node n_{EKG} in EKG and build a set PN_{EKG} of possible nodes. For each pair of nodes $(n_{EKG,i}, n_{EKG,j})$ in PN_{EKG} we find the shortest path sp_i from $n_{EKG,i}$ to $n_{EKG,j}$, if exists, and

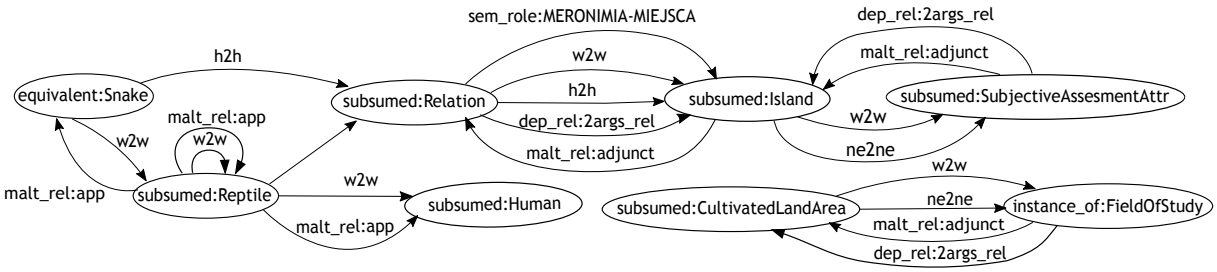

Figure 1: Graph built for sentence $S_{example}$ with *Concept* node type and full set of edges types.

add sp_i to temporary graph $G_{T(S(EKG))}$. After this process $G_{S(EKG)} = G_{T(S(EKG))}$. Using this procedure we can be built three merged graphs.

With plWordNet, $G_{merged} = G \cup EKG_{S(plwn)}$ includes nodes of the type synset (from the first step) edges built in *second step* and edges – relations from plWordNet subgraph.

With SUMO, $G_{merged} = G \cup EKG_{S(sumo)}$ includes concept nodes from the sentence and from the subgraph of SUMO Ontology. The edges are the relations from sentence and relations from the SUMO subgraph.

With plWordNet and SUMO, $G_{merged} = G \cup EKG_{S(plwn)} \cup EKG_{S(sumo)}$ contains full set of nodes: built in *first step*, from plWordNet and SUMO subgraphs, i.e. edges of all types.

There are 12 possible graph types in total, i.e. 4 types of nodes and 3 types of merge with both EKG, namely: *Lemma lower* graph merged with $EKG_{S(SUMO)}$, *Lemma PoS lower* merged with $EKG_{S(plwn)}$, *Concept* merged with $EKG_{S(sumo)}$ or *Synset* graph merged with $EKG_{S(plwn)} \cup EKG_{S(sumo)}$.

4.2.2 Similarity-based Features

For each instance of relation $R_i(S1, S2)$, a sentence pair, from the annotated corpus, see Sec. 3 16 graphs were built for both sentences $S1$ and $S2$: 4 graphs with different node types in the *second step* and 12 graphs with combinations of every node type with both EKG. Thus, each instance of relation R_i is assigned 16 graph-based representations of sentences $R_i(S1, S2) \Rightarrow R_{ik}(G1_k, G2_k), k \in < 1, \ldots 16 >$. Next, we calculate 8 different similarity measures between the graphs for R_i, including 7 similarity measures from the literature and one proposed by us. The measures are explained further on in this section. A single instance of relation R_i from the corpus is converted into a training vector v_i of the size 128 (16 graphs $\times$ 8 measures). The first measure is well known **Graph Edit Distance** (Fernández and Valiente, 2001) (GED), whose value is the minimal sum of the costs c (labelled as $\gamma(M)$) of atomic operations transforming G_1 to G_2:

$$GED(G_1, G_2) = min(\gamma(M)) \quad (1)$$

MCS (Bunke and Shearer, 1998) is the ratio of the size of *maximum common subgraph* (mcs) of G_1 and G_2 to the size of bigger graph of ($G1$ or $G2$):

$$MCS(G_1, G_2) = \frac{|mcs(G_1, G_2)|}{max\{|G_1|, |G_2|\}} \quad (2)$$

Measure **WGU** (Wallis et al., 2001) depends on calculating the ratio of the size of *mcs* G_1 and G_2 to the sum of sizes of both graphs minus *mcs* size:

$$WGU(G_1, G_2) = \frac{|mcs(G_1, G_2)|}{|G_1| + |G_2| - |mcs(G_1, G_2)|} \quad (3)$$

UGU (Bunke, 1997) is a simple measure, whose value is the difference between the sizes of G_1 and G_2 and the double size of of *mcs* G_1 and G_2:

$$UGU(G_1, G_2) = |G_1| + |G_2| - 2 \cdot |mcs(G_1, G_2)| \quad (4)$$

Next measure called **MMCS** was proposed by Fernández and Valiente (2001). The *MMCS* value expresses the dissimilarity of graphs G_1 and G_2:

$$MMCS(G_1, G_2) = |MCS(G_1, G_2)| - |mcs(G_1, G_2)| \quad (5)$$

Measure **MMCSN** (Fernández and Valiente, 2001) depends on calculating ratio of *mcs* and *MCS* for graphs G_1 and G_2.

$$MMCSN(G_1, G_2) = \frac{|mcs(G_1, G_2)|}{|MCS(G_1, G_2)|} \quad (6)$$

The last measure from literature is **Jaccard** similarity (Jaccard, 1912):

$$J(G_1, G_2) = \frac{|G_1 \cap G_2|}{|G_1 \cup G_2|} \quad (7)$$

We propose a simple extension of *Jaccard* measure, called **Contextual BOW**, Eq. (8). In it, the

context (neighborhood) of the node n_i from G_1 is compared with the context of node n_i in G_2. The neighborhood of node n in graph G is defined as input nodes $G(n)_{in}$ and output nodes $G(n)_{out}$.

$$N(G_1(n)) = \{G_1(n)_{in} \cup G_1(n)_{out}\}$$
$$N(G_2(n)) = \{G_2(n)_{in} \cup G_2(n)_{out}\}$$
$$S(N(G_1(n), G_2(n))) = \frac{|N(G_1(n)) \cap N(G_2(n))|}{|N(G_1(n)) \cup N(G_2(n))|}$$
$$G_{min} = G_1 \iff |G_1| \leq |G_2|$$
$$G_{min} = G_2 \iff |G_2| < |G_1|$$

Where $N(G_1(n))$ is the neighborhood of node n in G_1, and $N(G_2(n))$ of node n in G_2. The value of *CTXBowSim* is calculated as:

$$Sim(G_1, G_2) = CTXBowSim(G_1, G_2) = \frac{\sum_{n \in G_{min}}^{n} S(N(G_1(n), G_2(n)))}{|G_{min}|} \quad (8)$$

The similarity values are used as features during supervised learning to build a classifier. By changing the way of constructing the graphs and computing their similarity we tune the classification process into different aspects of the sentences being compared. The number of features generated for classification is dependent on the number of different graphs types, used to compare sentences, and the number of applied measures for calculating their similarity. Thus, it is a combination of all node representations, all EKG sources and the applied similarity measures.

5 Results and Evaluation

The corpus contains 3469 examples annotated with one of the possible CST relations. For classification we used SVM (*Suport Vectors Machine* (Steinwart and Christmann, 2008)) and LMT (*Logistic Model Tree* (Landwehr et al., 2005)). The classifiers were evaluated according to 10-fold cross-validation scheme (Kohavi, 1995).

First, the baseline set of features was tested, see Sec. 4.1. The classifiers were tested on relation types, which implies that the training set for the classification was highly unbalanced with respect to different relations. Table 1 shows the results for SVM and LMT and the baseline feature set. Zero values occurred for very specific relations with a small number of instances, e.g. 3 instances of Citation. Moreover, baseline features express only weak discrimination power.

In a multiclass setting, the average F-score value for SVM was 0.334 and 0.309 for LMT.

	SVM			LMT		
Rel.	P	R	F	P	R	F
Cita.	0.000	0.000	0.000	0.000	0.000	0.000
Foll.	0.583	0.023	0.044	0.000	0.000	0.000
Over.	0.454	0.985	0.622	0.465	0.967	0.628
Moda.	0.000	0.000	0.000	0.000	0.000	0.000
IS	0.000	0.000	0.000	0.000	0.000	0.000
Desc.	0.250	0.008	0.016	0.000	0.000	0.000
Equi.	0.000	0.000	0.000	0.000	0.000	0.000
Fulf.	0.000	0.000	0.000	0.000	0.000	0.000
Cont.	0.000	0.000	0.000	0.000	0.000	0.000
Sum.	0.000	0.000	0.000	0.000	0.000	0.000
HB	0.000	0.000	0.000	0.000	0.000	0.000
Iden.	0.900	0.150	0.257	0.430	0.767	0.551
Elab.	0.000	0.000	0.000	0.000	0.000	0.000
Subs.	0.429	0.031	0.058	0.492	0.160	0.241
Chan.	0.000	0.000	0.000	0.000	0.000	0.000
Sour.	0.000	0.000	0.000	0.000	0.000	0.000
NR	0.521	0.246	0.334	0.230	0.116	0.154
Avg.	0.349	0.457	0.307	0.254	0.457	0.309

Table 1: Results for the classifiers trained on the baseline feature set (lexical, syntactic, semantic).

Many CST relations were not recognized at all. Classifiers showed poor precision and recall in the relations detection task (*No relation* result), which means they could not decide whether a pair of sentences represents a CST link or not. The performance at recognition of relations was unsatisfactory, even for the most frequent relations including *Overlap*, *Follow-up*, *Subsumption* or *Description*.

For the graph-based approach, SVM and LMT were used again. Table 2 contains summarized results of classifiers trained with graph-based features. The performance achieved using graph-based features was better than in the previous approach. A significant improvement could be observed for both SVM and LMT. Only for the less frequent relations the classifiers were not able to correctly recognize the type. The average F-score value was 0.442 for SVM and 0.772 for LMT. We can note that LMT outperforms SVM in the classification on almost every class.

Table 3 shows the achieved results on a combined set of the baseline and graph-based features. A combination of these features had a positive impact on the performance of selected classifiers. The average F-score value was increased to 0.749 for SVM and 0.817 for LMT. Our method recognized even more complex relations like *Historical Background*, *Follow-up* or *Elaboration*, with good precision and slightly lower recall. Some of the relations that occur quite rarely in our dataset were also recognized, although performance for them was still low. The corpus used for evaluation has an irregular distribution of CST relations, nega-

	SVM			LMT		
Rel.	P	R	F	P	R	F
Cita.	0.000	0.000	0.000	1.000	0.333	0.500
Foll.	0.965	0.180	0.303	0.772	0.853	0.811
Over.	0.510	0.999	0.675	0.969	0.993	0.981
Moda.	0.000	0.000	0.000	0.000	0.000	0.000
IS	0.750	0.462	0.571	0.000	0.000	0.000
Desc.	0.578	0.070	0.125	0.556	0.739	0.634
Equi.	0.667	0.083	0.148	0.286	0.167	0.211
Fulf.	0.667	0.063	0.114	0.531	0.269	0.357
Cont.	0.000	0.000	0.000	0.000	0.000	0.000
Sum.	0.174	0.073	0.103	0.222	0.073	0.110
HB	0.727	0.103	0.180	0.643	0.756	0.695
Iden.	0.898	0.733	0.807	0.902	0.917	0.909
Elab.	0.378	0.114	0.175	0.707	0.431	0.535
Subs.	0.641	0.129	0.215	0.489	0.474	0.482
Chan.	0.000	0.000	0.000	0.000	0.000	0.000
Sour.	1.000	0.820	0.901	0.813	0.520	0.634
NR	0.956	0.437	0.600	0.776	0.749	0.762
Avg.	0.620	0.544	0.448	0.771	0.786	0.772

Table 2: The results for a graph-based approach.

tively affecting the results of classification. We can notice that for less frequent relations like *Citation*, *Modality*, *Indirect Speech* or *Contradiction*, the classifiers were not able to properly recognize types of the CST links.

	SVM			LMT		
Rel.	P	R	F	P	R	F
Cita.	0.000	0.000	0.000	0.000	0.000	0.000
Foll.	0.800	0.967	0.876	0.964	0.961	0.962
Over.	0.947	1.000	0.973	0.980	0.986	0.983
Moda.	0.000	0.000	0.000	0.000	0.000	0.000
IS	0.000	0.000	0.000	0.393	0.423	0.407
Desc.	0.551	0.728	0.627	0.613	0.707	0.657
Equi.	0.333	0.042	0.074	0.295	0.271	0.283
Fulf.	0.710	0.138	0.230	0.561	0.431	0.488
Cont.	0.000	0.000	0.000	0.167	0.150	0.158
Sum.	0.000	0.000	0.000	0.243	0.167	0.198
HB	0.565	0.724	0.635	0.695	0.753	0.723
Iden.	0.887	0.917	0.902	0.948	0.917	0.932
Elab.	0.933	0.341	0.500	0.607	0.577	0.592
Subs.	0.500	0.629	0.557	0.580	0.526	0.551
Chan.	0.000	0.000	0.000	0.000	0.000	0.000
Sour.	0.800	0.160	0.267	0.818	0.720	0.766
NR	0.777	0.723	0.749	0.873	0.868	0.871
Avg.	0.769	0.786	0.755	0.816	0.820	0.817

Table 3: The results for a combined approach - basis features extended with graph-based features.

As it was noted earlier, a similar distribution of the relations can be observed in the CSTNews corpus (Cardoso et al., 2011). The authors of CSTNews built it from news documents, i.e. the sources were very similar to those utilised in the corpus applied in this work. In (Maziero et al., 2014) CSTNews was used to evaluate recognition methods for the refined CST model. The authors stated that their classifier outperforms other CST parsers. Tab. 4 presents the results of our evaluation in comparison to the results reported in (Maziero et al., 2014). The comparison was indirect due to the different languages and data sets, but as both corpora have similar content and structure, this comparison can be informative.

	(Maziero et al., 2014)			Our LMT		
Rel.	P	R	F	P	R	F
Cita.	—	—	—	0.000	0.000	0.000
Foll.	0.282	0.273	0.277	0.964	0.961	0.962
Over.	0.441	0.478	0.458	0.980	0.986	0.983
Moda.	—	—	—	0.000	0.000	0.000
IS	0.529	0.632	0.576	0.393	0.423	0.407
Desc.	—	—	—	0.613	0.707	0.657
Equi.	0.378	0.359	0.368	0.295	0.271	0.283
Fulf.	—	—	—	0.561	0.431	0.488
Cont.	0.273	0.177	0.214	0.167	0.150	0.158
Sum.	—	—	—	0.243	0.167	0.198
HB	0.299	0.260	0.278	0.695	0.753	0.723
Iden.	1.000	1.000	1.000	0.948	0.917	0.932
Elab.	0.405	0.385	0.395	0.607	0.577	0.592
Subs.	0.449	0.447	0.448	0.580	0.526	0.551
Chan.	—	—	—	0.000	0.000	0.000
Sour.	—	—	—	0.818	0.720	0.766
NR	0.773	0.527	0.627	0.873	0.868	0.871
Tran.	0.500	0.500	0.500	—	—	—
Avg.	0.484	0.458	0.467	0.816	0.820	0.817

Table 4: Comparison of the results.

6 Conclusions

In our approach a sentence S is represented by different graphs referring to many types of the word-level representations. It is possible to express the same sentence S on the morphological level (*Lemma PoS Node type*) and/or semantic level (*Synset Node type*). By merging the graphs built from S with some external knowledge graph, we can expand the information stored in the graph of S and calculate similarity between graphs more accurately. The proposed approach to build graphs is language independent and is not depended on the existence of deeper parsers.

Relations extracted from sentence structures, i.e. *semantic roles* or *syntactic dependencies*, and lexical semantic representation assigned to words, i.e. *disambiguated senses* and *SUMO concepts*, were helpful in discriminating CST relation types. In our work we proposed a method for the recognition of the full set of 17 CST relations, in contrast to the limited of subsets used in literature, e.g. in (Kumar et al., 2012a). Our method outperforms also the state of the art algorithm when compared on a corpus of the similar origin and content.

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*. ACM, New York, NY, USA, WebMedia '08, pages 298–303.

Bartosz Broda, Paweł Kędzia, Michał Marcińczuk, Adam Radziszewski, Radosław Ramocki, and Adam Wardyński. 2013. *Fextor: A Feature Extraction Framework for Natural Language Processing: A Case Study in Word Sense Disambiguation, Relation Recognition and Anaphora Resolution*, Springer Berlin Heidelberg, Berlin, Heidelberg, pages 41–62.

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*. Cuiabá, Brazil, pages 88–105.

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.

Paul Jaccard. 1912. The Distribution of the Flora in the Alpine Zone . *New Phytologist* 11(2):37–50.

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

Ron Kohavi. 1995. A Study of Cross-validation and Bootstrap for Accuracy Estimation and Model Selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, IJCAI'95, pages 1137–1143.

Yogan Jaya Kumar, Naomie Salim, Albaraa Abuobieda, and Ameer Tawfik Albaham. 2014. Multi document summarization based on news components using fuzzy cross-document relations. *Applied Soft Computing* 21:265–279.

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. 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):161–205.

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

Marek Maziarz, Maciej Piasecki, Ewa Rudnicka, Stan Szpakowicz, and Pawel Kedzia. 2016. plWordNet 3.0 - a Comprehensive Lexical-Semantic Resource. In *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan*. pages 2259–2268.

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.

Joakim Nivre, Johan Hall, Jens Nilsson, Atanas Chanev, Gülsen Eryigit, Sandra Kübler, Svetoslav Marinov, and Erwin Marsi. 2007. MaltParser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering* 13(02):95–135.

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.

Maciej Piasecki, Stanisław Szpakowicz, and Bartosz Broda. 2009. *A Wordnet from the Ground Up*. Oficyna Wydawnicza Politechniki Wroclawskiej, Wrocław.

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*. Association for Computational Linguistics, Stroudsburg, PA, USA, SIGDIAL '00, pages 74–83.

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. In H. Rybiński M. Kryszkiewicz M. Niezgódka R. Bembenik, Ł. Skonieczny, editor, *Intelligent Tools for Building a Scientific Information Platform: Advanced Architectures and Solutions*, Springer Verlag, page to appear.

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

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.

Ingo Steinwart and Andreas Christmann. 2008. *Support Vector Machines*. Springer Publishing Company, Incorporated, 1st edition.

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

Marcin Woliński. 2006. Morfeusz — a practical tool for the morphological analysis of Polish. In Mieczysław A. Kłopotek, Sławomir T. Wierzchoń, and Krzysztof Trojanowski, editors, *Intelligent Information Processing and Web Mining*, Springer-Verlag, Berlin, Advances in Soft Computing, pages 503–512.

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

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*. ACM, New York, NY, USA, CIKM '03, pages 124–130.

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*. Springer-Verlag, Berlin, Heidelberg, IJCNLP'04, pages 32–41.