K-Best Spanning Tree Dependency Parsing With Verb Valency Lexicon Reranking

Željko AGIĆ
Department of Information and Communication Sciences
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10000 Zagreb, Croatia
zeljko.agic@ffzg.hr

ABSTRACT
A novel method for hybrid graph-based dependency parsing of natural language text is proposed. It is based on k-best maximum spanning tree dependency parsing and evaluation of the spanning trees by using a verb valency lexicon for a given language as a reranking knowledge base. The approach is compared with existing state-of-the-art transition-based and graph-based approaches to dependency parsing. As the proposed generic method was developed specifically for improving the accuracy of Croatian dependency parsing, Croatian Dependency Treebank and CROVALLEX verb valency lexicon are used in the experiment. The suggested approach scored approximately 77.21% LAS, outperforming the tested state-of-the-art approaches by at least 2.68% LAS.

TITLE AND ABSTRACT IN CROATIAN

Ovisnosno parsanje pomoću *k* najboljih razapinjućih stabala i ponovnoga vrjednovanja valencijskim rječnikom glagola

Predlaže se novi pristup hibridnom ovisnosnom parsanju tekstova prirodnoga jezika temeljenom na teoriji grafova. Pristup je zasnovan na ovisnosnom parsanju pomoću *k* najboljih razapinjućih stabala i uporabi valencijskog rječnika glagola parsanoga jezika kao baze znanja za ponovno vrjednovanje tih stabala. Pristup je uspoređen s najboljim postojećim pristupima ovisnosnom parsanju temeljenima na teoriji grafova i na prijelazničkim sustavima. Budući da je predložena metoda razvijana sa specifičnim ciljem povećanja točnosti ovisnosnoga parsanja hrvatskih tekstova, u eksperimentu je korištena Hrvatska ovisnosna banka stabala i valencijski rječnik glagola hrvatskoga jezika CROVALLEX. Predloženi pristup postigao je ukupnu točnost od otprilike 77.21% LAS, što predstavlja povećanje točnosti od oko 2.68% LAS u odnosu na testirane najbolje postojeće sustave.

KEYWORDS: dependency parsing, k-best spanning trees, verb valency lexicon.

KEYWORDS IN CROATIAN: ovisnosno parsanje, razapinjuća stabla, valencijski rječnik glagola.

Proceedings of COLING 2012: Posters, pages 1–12,
COLING 2012, Mumbai, December 2012.

1 Condensed version of the paper in Croatian

Kvaliteta parsanja u paradigmi ovisnosnoga parsanja temeljenog na podatcima ovisi u najvećoj mjeri o svojstvima parsanoga jezika. Budući da su svojstva jezika u tome teorijskom okviru implicitno sadržana u banci ovisnosnih stabala, kaže se da je kvaliteta parsanja ovisna o svojstvima banke ovisnosnih stabala (Kübler et al., 2009). Ovdje se nastoji — koristeći postojeće spoznaje o ovisnosnomu parsanju različitih razreda prirodnih jezika parserima temeljenim na podatcima (Buchholz and Marsi, 2006; Nivre et al., 2007) — unaprijediti kvalitetu ovisnosnoga parsanja tekstova pisanih hrvatskim jezikom koristeći postojeće metode ovisnosnoga parsanja, Hrvatsku ovisnosnu banku stabala (HOBS) (Tadić, 2007) i valencijski rječnik glagola hrvatskoga jezika CROVALLEX (Mikelić Preradović, 2008; Mikelić Preradović et al., 2009).

Prikazana su dva skupa eksperimenata. U prvome se skupu na hrvatskim tekstovima iz HOBS-a testiraju najbolji od postojećih javno dostupnih ovisnosnih parsera temeljenih na podatcima, kako bi se utvrdila najveća točnost parsanja koja se može postići njihovom uporabom. S obzirom na točnost postignutu pri parsanju srodnih jezika, poput češkoga i slovenskoga, u sklopu natjecanja u ovisnosnome parsanju CoNLL 2006 (Buchholz and Marsi, 2006) i 2007 (Nivre et al., 2007), za vrjednovanje je odabran MaltParser (Nivre et al., 2007) kao najbolji predstavnik prijelazničkih parsera i MSTParser (McDonald et al., 2006) kao najbolji među ovisnosnim parserima temeljenima na teoriji grafova. Za testiranje je korištena najnovija inačica HOBS-a, koja je sadržavala ukupno 88,045 pojavnica u 3,465 rečenica. Osnovni statistički podatci o HOBS-u i skupovima za treniranje i testiranje ovisnosnih parsera izloženi su u Tablici 1. Sva mjerenja su ponovljena deset puta, i to podjelom HOBS-a na deset nepreklapajućih dijelova, korištenjem devet od tih deset dijelova za postupak treniranja i desetoga dijela, veličine oko 5,000 pojavnica, za postupak testiranja. Uporabljeno je sedam algoritama za prijelazničko parsanje iz sustava MaltParser i četiri algoritma za parsanje temeljeno na teoriji grafova iz sustava MSTParser. Slika 1 i Tablica 2 i 3 prikazuju rezultate prvoga skupa eksperimenata. Parser MstCle2 (neprojektivni ovisnosni parser temeljen na grafovima, jezičnome modelu s parovima ovisnosnih relacija i algoritmu za pronalaženje najvećega prostirućeg stabla Chu-Liu/Edmonds) postigao je najveću točnost pri parsanju tekstova iz HOBS-a prema svim odabranim mjerama za vrjednovanje. Testiranje statističke značajnosti pokazalo je da su razlike u točnostima svih parsera temeljenih na grafovima u odnosu na prijelazničke parsere statistički značajne. S druge

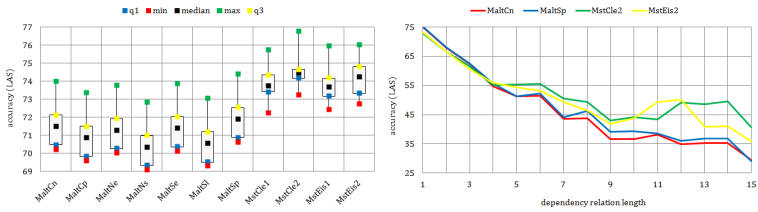

Figure 1: Overall parsing accuracy boxplot and parsing accuracy with respect to dependency relation length for the top-performing algorithms of the standard dependency parsers (Croatian: Ukupna točnost parsanja algoritmima postojećih ovisnosnih parsera i točnost parsanja s obzirom na duljinu ovisnosne relacije za najbolje od tih algoritama)

Figure 2: Two CROVALLEX valency frames for the verb *dotaknuti* (en. *to touch*) (Croatian: Dva valencijska okvira glagola *dotaknuti* u CROVALLEX-u)

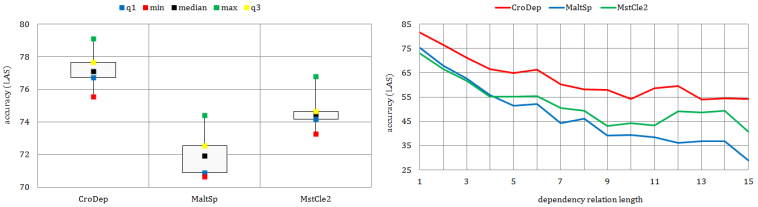

Figure 3: Overall parsing accuracy boxplot and parsing accuracy with respect to dependency relation length for CroDep, MaltSp and MstCle2 algorithm (Croatian: Ukupna točnost parsanja i točnost parsanja s obzirom na duljinu ovisnosne relacije za algoritme CroDep, MaltSp i MstCle2)

strane, točnosti među parserima unutar skupine prijelazničkih parsera nisu različite u statistički značajnoj mjeri. Razlike u postignutoj točnosti parsanja nisu statistički značajne ni u skupini parsera temeljenih na grafovima, no u njoj se po postignutoj točnosti izdvajaju parseri MstCle2 (74.53% LAS) i MstEis2 (74.17% LAS), odnosno parseri s jezičnim modelima temeljenim na parovima ovisnosnih relacija.

U drugome se skupu eksperimenata uspoređuje parser CroDep s najboljim parserima iz prvoga skupa eksperimenata. Parser CroDep (Agić, 2012) novopredloženi je hibridni ovisnosni parser koji se sastoji od tri međuovisne komponente: (1) ovisnosnoga parsera temeljenog na teoriji grafova u skladu s izvedbom iz (McDonald et al., 2005) i algoritmu za parsanje pronalaženjem *k* najboljih parsanja ulazne rečenice u skladu s izvedbom iz (Hall, 2007), (2) vrjednovatelja predloženih *k* ovisnosnih stabala pomoću valencijskoga rječnika glagola hrvatskoga jezika CROVALLEX i (3) modula za ponovno vrjednovanje tih stabala povezivanjem vrjednovanja iz dviju prethodnih komponenata, koji daje konačni izlaz iz sustava u vidu jednoga ovisnosnog stabla kojem je dodijeljena najviša zbirna ocjena. Slika 3 i Tablica 5 i 6 prikazuju rezultate drugoga skupa eksperimenata. Zabilježena je točnost parsera CroDep od 77.21% prema mjeri LAS, što predstavlja porast od 2.68% u odnosu najbolji parser iz prethodnoga skupa eksperimenata. Razlika između njihovih točnosti statistički je značajna s obzirom na sve korištene mjere. Za porast ukupne točnosti zaslužan je statistički značajan porast točnosti parsanja imenica i glagola. Prema mjerama LAS i UAS parser CroDep u usporedbi s parserom MstCle2 bilježi povećanje točnosti od preko 10% za predikate, subjekte i objekte, što potvrđuje smislenost povezivanja CROVALLEX-a i parsera temeljenoga na grafovima. Detaljniji prikaz izvedbe parsera CroDep i rezultata pojedinih ekperimenata izložen je dalje u tekstu.

2 Introduction

The quality of data-driven dependency parsing — as expressed by the *de facto* standard dependency parsing evaluation metrics such as LAS and UAS (Nivre, 2006) — is repeatedly shown to be highly language-dependent. More specifically, being that syntactic properties of a given language are implicitly encoded by dependency treebanks in the framework of data-driven dependency parsing, it is seen as treebank-dependent (Kübler et al., 2009). The CoNLL 2006 and 2007 shared tasks on multilingual data-driven dependency parsing (Buchholz and Marsi, 2006; Nivre et al., 2007) separate the tested languages into three classes on the basis of observed dependency parsing accuracy scores: low, medium and high. It is specifically noted in (Nivre et al., 2007) that "the classes are more easily definable via language characteristics than via characteristics of the data sets" and that the "most difficult[-to-parse] languages are those that combine a relatively free word order with a high degree of inflection," modified to an extent only by the respective treebank sizes.

The research presented here was conducted with a goal of improving the baseline for dependency parsing of Croatian texts. Croatian is a highly inflected Slavic language with a relatively free word order, similar to Czech and Slovene, which were included in the CoNLL shared tasks on dependency parsing. With respect to the results of the shared tasks, it was expected that the scores for state-of-the-art parsers on the Croatian Dependency Treebank (Tadić, 2007; Agić, 2012; Berović et al., 2012) would place Croatian in the low accuracy class, i.e., the class of most difficult-to-parse languages. After conducting these preliminary experiments, various courses of action were considered in order to improve the baseline. Parsing baselines in data-driven dependency parsing are usually topped by feature reselection (Passarotti and Dell'Orletta, 2010), merging parser outputs — as in parser voting (Sagae and Lavie, 2006) and stacking (Nivre and McDonald, 2008) — and by using external sources of linguistic knowledge, such as subcategorization information (Zeman, 2002), possibly introducing rule-based (language-dependent) modules into data-driven (language-independent) parsers. The presented approach deals with implementing an interaction between a graph-based dependency parser and a valency lexicon of Croatian verbs, producing in turn a parsing model requiring only a dependency treebank and a machine-readable verb valency lexicon to operate.

In the paper, the baseline experiment in dependency parsing of Croatian using existing state-of-the-art data-driven dependency parsers is first described, including a description of the current state of development of the Croatian Dependency Treebank. Second, the valency lexicon reranking parser is introduced, along with a short description of the verb valency lexicon of Croatian verbs used in this experiment — CROVALLEX (Mikelić Preradović, 2008; Mikelić Preradović et al., 2009). The newly-developed parser is evaluated within the testing framework of the existing parsers and the obtained results are presented in comparison. Future work plans for improving the parser and for improving dependency parsing of Croatian texts in general are sketched in the closing section.

3 Baseline experiment

At the time of conducting the experiments within the CoNLL shared tasks of 2006 and 2007, no treebanks for Croatian were available, for any syntactic formalism. More precisely, the development of the Croatian Dependency Treebank started in January 2007 and only a 100-sentence prototype of the treebank existed when the CoNLL 2007 task was initiated. Being that both shared tasks required a training and testing set and that the minimum size of the testing set was fixed at 5,000 tokens, the prototype was insufficient for participation. Once

Feature	Entire treebank	Training sets	Testing sets
Sentences	3,465	3,261.18 ± 4.20	203.82 ± 4.20
Tokens	88,045	82,865.88 ± 6.87	5,179.12 ± 6.87
Types (wordforms)	20,703	19,927.06 ± 15.71	2,594.06 ± 12.26
Lemmas	10,481	10,166.00 ± 9.19	1,909.00 ± 14.12
POS/MSD tags	828	817.94 ± 1.40	368.35 ± 4.41
Analytical functions	26	26.00 ± 0.00	23.24 ± 0.43

Table 1: Basic stats for Croatian Dependency Treebank and its tenfold sets

the treebank had finally matured in size, the shared task experiments could be recreated for Croatian texts in order to establish a baseline. This section briefly presents this experiment by presenting the treebank, parser selection, experimental setup and the obtained results.

3.1 Treebank

Quoting (Tadić, 2007) and (Agić et al., 2010), the Croatian Dependency Treebank (hr. *Hrvatska ovisnosna banka stabala*, HOBS further in the text) is a dependency treebank built along the principles of Functional Generative Description (Sgall et al., 1986), a multistratal model of dependency grammar developed for Czech. This formalism was further adapted in the Prague Dependency Treebank (PDT) (Hajič et al., 2000) project and applied in sentence analysis and annotation on the levels of morphology, syntax and tectogrammatics. The ongoing construction of HOBS closely follows the guidelines set by PDT, with their simultaneous adaptation to the specifics of Croatian. More detailed account of the HOBS project plan is given in (Tadić, 2007). Currently, HOBS consists of 3,465 sentences in the form of dependency trees that were manually annotated with syntactic functions using TrEd (Pajas, 2000) as the annotation tool. These sentences, encompassing approximately 88,000 tokens, stem from the CW100 newspaper sub-corpus of the Croatian National Corpus (Tadić, 2002, 2009). The CW100 sub-corpus was previously XCES-encoded, sentence-delimited, tokenized, lemmatized and morphosyntactically annotated by linguists. Thus, each of the analyzed sentences contains the manually assigned information on part-of-speech, morphosyntactic category, lemma, dependency and analytical function for each of the tokens. Such a course of action was taken in order to enable the training procedures of state-of-the-art dependency parsers to choose from a wide selection of different features in experiments with stochastic dependency parsing of Croatian texts. Basic stats for HOBS and the experiment sets are given in Table 1. Sentences in HOBS are annotated according to the PDT annotation manual for the analytical level of annotation, with respect to differing properties of Croatian and consulting the Slovene Dependency Treebank (SDT) project (Džeroski et al., 2006). The utilized analytical functions are thus compatible with those of PDT. HOBS is available via META-SHARE (Federmann et al., 2012; Piperidis, 2012).

The experiment was envisioned as a tenfold cross-validated run of several parsing algorithms on the Croatian Dependency Treebank respecting the rules of the CoNLL 2006 and 2007 shared tasks. Training and testing set stats are also given in Table 1. They are indicative of the high morphological complexity of Croatian, as tokens in HOBS are annotated by using 828 different morphosnytactic tags (out of the 1,405 existing in the Croatian Morphological Lexicon (Tadić and Fulgosi, 2003)). As to the syntactic complexity inherent in HOBS, 1,801 of dependency relations (2.06%) were found to be non-projective in 761 different sentences

System	Algorithm	LA	LAS	UAS
MaltParser	Nivre eager	83.74 ± 0.46	71.29 ± 0.74	77.13 ± 0.71
	Nivre standard	83.16 ± 0.47	70.35 ± 0.73	76.44 ± 0.70
	Covington projective	83.46 ± 0.48	70.87 ± 0.73	76.80 ± 0.69
	Stack projective	84.05 ± 0.44	71.91 ± 0.74	77.59 ± 0.73
MaltParser	**Covington non-projective**	83.88 ± 0.46	71.50 ± 0.74	77.30 ± 0.72
	Stack eager	83.75 ± 0.42	71.39 ± 0.73	77.23 ± 0.72
	Stack lazy	83.28 ± 0.48	70.56 ± 0.73	76.54 ± 0.71
MSTParser	Eisner 1st	85.57 ± 0.36	73.73 ± 0.65	80.92 ± 0.61
	Eisner 2nd	85.64 ± 0.39	74.17 ± 0.64	81.27 ± 0.59
MSTParser	Chu-Liu/Edmonds 1st	85.76 ± 0.35	73.88 ± 0.58	80.99 ± 0.50
	Chu-Liu/Edmonds 2nd	85.87 ± 0.38	74.53 ± 0.57	81.69 ± 0.44

Table 2: Overall parsing accuracy of the standard dependency parsing algorithms

(21.96%), indicating an expectedly high presence of non-projectivity, similar to what is observed in PDT (Nivre and Nilsson, 2005). All selected parsers were thus evaluated as non-projective parsers, regardless of their need for treebank (de)projectivization that may be present as a pre- or post-processing step in certain workflows.

3.2 Parsers

Parser selection was based on the results of the CoNLL 2006 and 2007 shared tasks for languages similar to Croatian, i.e., the observed LAS scores for Czech and especially Slovene (being that PDT is substantially larger than both HOBS and SDT). Two standalone parser generators — MaltParser (Nivre et al., 2007) and MSTParser (McDonald et al., 2006) — were shown to be predominant in scores for parsing morphologically complex languages, with the graph-based MSTParser systems slightly outperforming the transition-based MaltParser systems for both emphasized languages. Based on these results, MaltParser and MSTParser were chosen among the publicly available CoNLL 2006 and 2007 parser generators for inclusion in the baseline testing on HOBS. MaltParser was configured by using MaltOptimizer (Ballesteros and Nivre, 2012) and seven projective and non-projective parsing algorithms were tested, while four different configurations of MSTParser were tested: first- and second-order arc-factored language model with Eisner's (projective) and Chu-Liu/Edmonds (non-projective) parsing algorithms.

3.3 Results

Labeled (LAS) and unlabeled (UAS) attachment score was observed, as well as linear label attachment (LA), both overall and for specific syntactic functions and parts of speech. The first set of results is given in Table 2. Systems are grouped into four classes by the way parsing algorithms handle non-projectivity — as pseudo-projective and non-projective MaltParsers and projective and non-projective MSTParsers. Boldfaced names indicate the top performing algorithms for the four classes. Statistical significance of the observed scores is indicated by Figure 1 as it shows that the MSTParser systems are consistently and significantly outperforming MaltParser systems. The top-scorer of the experiment is the MSTParser system that used a second-order arc-factored language model and the non-projective Chu-Liu/Edmonds maximum

Algorithm	Adv	Atr	AuxC	AuxP	Coord	Obj	Pnom	Pred	Sb
MaltSp	70.67	83.77	74.36	71.99	46.28	67.40	66.55	36.45	69.14
MaltCn	71.31	83.98	75.68	72.08	46.96	68.15	66.35	37.33	70.12
MstEis2	69.01	81.80	71.94	74.35	56.49	69.38	65.18	68.10	72.51
MstCle2	68.38	81.46	73.21	74.15	55.05	68.29	62.47	69.09	72.63

Table 3: Accuracy of the top-performing standard dependency parsing algorithms for specific syntactic functions

spanning tree (MST) algorithm. The results are consistent with the ones recorded for Czech and Slovene in CoNLL 2006 and 2007. Figure 1 also shows how graph-based top-performing systems handle long-range dependencies better than their transition-based counterparts.

From another perspective, the top-performing MSTParser system with a second-order arc-factored language model and the Chu-Liu/Edmonds MST parsing algorithm scored approximately 74.53% LAS, a result that places Croatian in the group of languages with low parsing accuracy, as expected. Table 3 additionally indicates that the parsing scores are even lower for the most important syntactic functions with respect to information extraction – subjects (Sb), predicates (Pred) and objects (Obj). This supported the initial estimation that a compound approach to dependency parsing should be implemented in order to increase overall parsing accuracy for Croatian. The method is presented and evaluated in the following section.

4 Valency lexicon reranking parser

The suggested parsing method draws on the fact that verb valency lexicons, such as VALLEX (Lopatková et al., 2006) and the VALLEX-inspired valency lexicon of Croatian verbs CROVALLEX (Mikelić Preradović et al., 2009), explicitly encode obligatory and optional constraints on the number and morphosyntactic properties of dependents that the verbs contained in the lexicon impose. While rule-based dependency parsers might use such information on (predicate) verbs at parser runtime, a post-processing reranking approach is presented here. Namely, a parser is developed that provides k-best dependency trees sorted by adequacy for an input sentence and this parser is then linked to the valency lexicon through a reranking module that reorders the suggested k-best trees by using the lexicon to evaluate them. The evaluation and subsequent reranking is done by weighting dependency relations whose heads are verbs contained within the valency lexicon and adding up the weights to provide overall scores for the suggested dependency trees.

In this section, the CROVALLEX valency lexicon is briefly presented and followed by a presentation of the reranking parser CroDep. The parser is then evaluated in the same testing environment as in the baseline experiment and the obtained results are discussed.

4.1 Valency lexicon

CROVALLEX is a verb valency lexicon created following the FGD guidelines (Sgall et al., 1986) and is accordingly compatible with HOBS with respect to the syntactic theory of choice. The utilized version of CROVALLEX (v2.008) contained 1,797 verb lemmas with 5,188 valency frames. Each valency frame is defined by stating the number, obligatoriness and morphological properties of sentence elements that a given verb introduces. An example is given in Figure 2

Sb	AuxP	AuxV	Obj	Adv	AuxC	Pnom
19.87%	16.38%	15.47%	12.17%	10.00%	5.34%	4.27%
Coord	AuxR	AuxY	AuxX	AuxT	AuxG	Apos
3.93%	2.01%	2.00%	2.00%	1.61%	1.42%	1.19%
AtvV	Pred	ExD	AuxZ	AuxK	AuxO	Atr
0.82%	0.65%	0.40%	0.35%	0.05%	0.05%	0.03%

Table 4: Distribution of direct predicate dependents in HOBS

for the verb *dotaknuti* (en. *to touch*). In the first frame, the agent (AGT) is obligatory (obl) and the instrument (INST) is optional and has to be in the instrumental case (case number 7). In the second frame, the patient (PAT) is obligatory and has to be in the accusative case (4).

The lexicon was adapted to the requirements of data-driven dependency parsing by filtering all multiword lemmas and entries with frequency of 0 denoted in the lexicon. 1,455 verbs and 4,090 respective frames were held. HOBS was then tested using CROVALLEX in order to observe the overlaps. HOBS contained a total of 1,525 verb lemmas and 12,952 verb tokens (ca 14.55% of all lemmas and 14.72% of all tokens), out of which a total of 791 verb lemmas was found in CROVALLEX (ca 51.87%). On the other hand, 664 of the CROVALLEX verbs were not represented by HOBS (ca 45.64% of all CROVALLEX verbs). Even though the measurement of static coverage of verb lemmas itself implicitly supports the course of action with interacting CROVALLEX with a HOBS-trained dependency parser, the dynamic coverage, i.e., the coverage of verb tokens provides an even stronger justification. Only 9.24% of all the verb tokens in HOBS were not covered by CROVALLEX, i.e., for 90.76% of verb tokens in HOBS, at least a single valency frame was found in CROVALLEX.

Another CROVALLEX-related viewpoint on HOBS is given in Table 4. It shows the relative frequency of dependents attached to verbal predicates by their syntactic function. It can be seen that a place in the sentence is most frequently opened by predicates to subjects (almost 20%), prepositions introducing prepositional phrases (16.38%), auxiliary verbs (15.47%), objects (12.17%) and adverbials (10%). The distribution indicates that the properties of verbs encoded in CROVALLEX are readily instantiated in HOBS.

4.2 Parser

Within the suggested framework, parsing is envisioned as a three-step procedure. First, k-best dependency trees sorted by confidence are provided by a language-independent data-driven parsing algorithm. Second, these k dependency trees are scored by a valency-lexicon-based scoring module. Third and final, the trees are re-sorted by combining the scores from the previous steps.

The data-driven component is a dependency parser based on both MSTParser (McDonald et al., 2006) and kMSTParser (Hall, 2007). Graph-based dependency parsing was chosen as a starting point in prototype development on basis of the results obtained in the baseline experiment, showing that graph-based dependency parsers consistently outperform transition-based dependency parsers on Croatian texts. The prototype uses the perceptron training algorithm implemented in MSTParser (McDonald et al., 2005) and the parsing algorithm based on (Camerini et al., 1980) for detecting k-best maximum spanning trees adapted to dependency

parsing in kMSTParser. This prototype parser is called CroDep0. Currently it supports only first-order arc-factored language models. It was evaluated on HOBS within the baseline testing framework to provide a reference point and it scored 73.27% LAS, 1.26% lower than the top-performing second-order Chu-Liu/Edmonds MSTParser.

The verb valency lexicon reranking component prototype was developed in what could be considered the simplest possible form of validating dependency relations with respect to valency frames. Namely, the reranking component takes a dependency tree as input. It searches the tree for verbal predicates. When a verbal predicate is encountered, its lemma is matched with the valency lexicon. If it exists as an entry in the lexicon, each of the first-level dependents introduced to the sentence by the verbal predicate is matched with the predicted slots in the valency frames on basis of its morphosyntactic properties: if the properties match and if the element is defined as obligatory, the tree score is incremented. The final score of the tree is defined as the sum of scores of all dependency relations having a verbal predicate as relation head. Each of the k-best trees provided by CroDep0 is given a score by the reranking component.

Finally, the re-sorting component combines the two lists — confidence scores for the k-best trees provided by CroDep0 and valency scores provided by the valency reranker — into a single list averaging the scores while favoring the stochastic component in case of ties. Being that the dependency tree scores from the valency reranker are positive integers representing overall counts of dependency relation confirmations extracted from the valency lexicon, they are normalized for comparison with the CroDep0 confidence scores. The normalization is done by using the maximum confidence score of CroDep0 as ceiling for the valency reranker scores. More formally, let $S_p = \{c_p(t_i)\}_{i=1}^{k}, \forall i, c_p(t_i) \in [0,1]$ represent the confidence scores for the k dependency trees t_i from the k-best parser module and let $S_v = \{c_v(t_i)\}_{i=1}^{k}, \forall i, c_v(t_i) \in \mathbb{N}$ be a list of trees and respective integer scores obtained from the valency reranker. The normalized valency reranker scores $\hat{S}_v$ and finally the overall dependency tree scores S_o are provided by the re-sorting component as follows.

$$\hat{S}_v = \{\hat{c}_v(t_i)\}_{i=1}^{k}, \forall i, \hat{c}_v(t_i) \in [0,1] \qquad \hat{c}_v(t_i) = \max_i(c_p(t_i)) \cdot \frac{c_v(t_i)}{\max_i(c_v(t_i))}$$

$$S_o = \{c_o(t_i)\}_{i=1}^{k}, \forall i, c_o(t_i) \in [0,1] \qquad c_o(t_i) = \frac{2 \cdot c_p(t_i) \cdot \hat{c}_v(t_i)}{c_p(t_i) + \hat{c}_v(t_i)}$$

The final output of the parser is always $\arg\max_i(c_o(t_i))$. If there are multiple dependency trees with the same overall score c_o, the ordering is decided by selecting the tree with the highest relative score in the k-best parser ranking, i.e., S_p. The resulting parser prototype is called CroDep. It inherits the properties of CroDep0 stochastic module, has the value of k fixed to 10 and additionally requires a verb valency lexicon in VALLEX-XML format for operation.

4.3 Results

Table 5 shows the overall accuracy of CroDep and its accuracy on selected parts of speech. CroDep outperforms the top-performing baseline parser by 2.68% LAS and the difference is shown to be statistically significant. The difference is also indicated graphically by the confidence intervals in Figure 3 (left side), where CroDep is compared to the top-performing graph-based system (second-order Chu-Liu/Edmonds MSTParser) and the top-performing transition-based

metric	Noun	Verb	Adj	Adp	Pro	Adv	overall
LA	85.34	87.89	92.67	98.64	84.38	80.14	88.27 ± 0.30
LAS	80.10	82.85	86.40	71.20	76.04	65.77	77.21 ± 0.59
UAS	90.16	86.84	89.13	71.92	84.84	75.30	83.05 ± 0.50

Table 5: Overall accuracy and accuracy on specific parts of speech for CroDep

metric	Adv	Atr	AuxC	AuxP	Coord	Obj	Pnom	Pred	Sb
LAS	70.69	83.94	69.80	70.59	49.41	83.17	71.46	82.12	85.01
UAS	84.81	88.90	71.53	71.48	50.87	93.12	79.92	86.81	91.35
P(LA)	78.96	91.21	91.96	97.86	89.72	84.12	77.06	84.36	86.78
R(LA)	74.11	90.94	87.77	97.74	81.60	94.75	49.73	97.21	97.50

Table 6: CroDep accuracy on specific syntactic functions

system (MaltParser stack projective). Table 6 shows the CroDep LAS and UAS scores on selected syntactic functions, as well as precision and recall with respect to label attachment for these functions, similar to linear morphosyntactic tagging evaluation. Compared to Table 3 which listed the scores on these syntactic functions for the top-performing baseline parsers, it can be clearly seen that the overall increase of CroDep accuracy by 2.68% LAS on MSTParser is caused by a substantial increase in LAS for predicates, subjects and objects (more than 10.00% LAS for each of the functions). Figure 3 (right side) shows that CroDep also handles long-distance dependencies better than the best baseline parsers and that its footprint is very similar to the one of the graph-based parser.

Conclusion and perspectives

A method is presented for hybrid language-independent dependency parsing by combining data-driven k-best maximum spanning tree parsing and rule-based reranking guided by a verb valency lexicon. It was tested in the form of prototype parser CroDep on the Croatian Dependency Treebank by using the CROVALLEX lexicon of Croatian verbs and it scored 77.21% LAS, topping the top-performing baseline parser by 2.68% LAS. Future work plans include testing the method on other languages, combining CroDep with other parsers and using methods of automatic valency frame extraction to enrich existing resources. Introduction of valency features to standard parsers might be considered. A preliminary experiment with parsing Czech was conducted by using PDT and VALLEX in compliance with the CoNLL 2007 shared task. CroDep scored 80.51% LAS, topping CroDep0 by 1.73% LAS, thus indicating method applicability across languages and outlining the influence of resource properties on method performance. Further research in language-independent k-best spanning tree parsing with valency lexicon reranking is required to support these preliminary results.

Acknowledgments

The presented results were partially obtained from research within project CESAR (ICT-PSP, grant 271022) funded by the European Commission, and partially from research within projects 130-1300646-0645 and 130-1300646-1776 funded by the Ministry of Science, Education and Sports of the Republic of Croatia.

References

Agić Ž, Šojat K, Tadić M. (2010). An Experiment in Verb Valency Frame Extraction from Croatian Dependency Treebank. In *Proceedings of ITI 2010*, Zagreb, SRCE University Computer Centre, University of Zagreb, 2010, pp. 55–60.

Agić Ž. (2012). *Pristupi ovisnosnom parsanju hrvatskih tekstova*. PhD thesis, University of Zagreb, Faculty of Humanities and Social Sciences, 2012.

Ballesteros M, Nivre J. (2012). MaltOptimizer: A System for MaltParser Optimization. In *Proceedings of LREC 2012*, ELRA, 2012.

Berović D, Agić Ž, Tadić M. (2012). Croatian Dependency Treebank: Recent Development and Initial Experiments. In *Proceedings of LREC 2012*, ELRA, 2012, pp. 1902–1906.

Buchholz S, Marsi E. (2006). CoNLL-X Shared Task on Multilingual Dependency Parsing. In *Proceedings of the 10th Conference on Computational Natural Language Learning (CoNLL-X)*, New York, NY, pp. 149–164.

Camerini P M, Fratta L, Maffioli F. (1980). The k-Best Spanning Arborescences of a Network. *Networks*, 10, 1980, pp. 91–110.

Džeroski S, Erjavec T, Ledinek N, Pajas P, Žabokrtský Z, Žele A. (2006). Towards a Slovene Dependency Treebank. In *Proceedings of LREC 2006*, ELRA, 2006.

Federmann C, Giannopoulou I, Girardi C, Hamon O, Mavroeidis D, Minutoli S, Schröder M. (2012). META-SHARE v2: An Open Network of Repositories for Language Resources including Data and Tools. In *Proceedings of LREC 2012*, ELRA, 2012, pp. 3300–3303. See URL http://www.meta-share.eu (accessed 2012-10-27).

Hajič J, Böhmová A, Hajičová E, Vidová Hladká B. (2000). The Prague Dependency Treebank: A Three-Level Annotation Scenario. In *Treebanks: Building and Using Parsed Corpora*, Amsterdam, Kluwer, 2000.

Hall K. (2007). K-Best Spanning Tree Parsing. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, 2007, pp. 392–399.

Kübler S, McDonald R, Nivre J. (2009). *Dependency Parsing*. Synthesis Lectures on Human Language Technologies, Morgan&Claypool Publishers, 2009.

Lopatková M, Žabokrtský Z, Skwarska K. (2006). Valency Lexicon of Czech Verbs: Alternation-Based Model. In *Proceedings of LREC 2006*, pp. 1728–1733.

McDonald R, Crammer K, Pereira F. (2005). Online Large-Margin Training of Dependency Parsers. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, ACL, 2005.

McDonald R, Lerman K, Pereira F. (2006). Multilingual Dependency Parsing with a Two-Stage Discriminative Parser. In *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X)*, 2006.

Mikelić Preradović N. (2008). *Pristupi izradi strojnog tezaurusa za hrvatski jezik*. PhD thesis, University of Zagreb, Faculty of Humanities and Social Sciences, 2008.

Mikelić Preradović N, Boras D, Kišiček S. (2009). CROVALLEX: Croatian Verb Valence Lexicon. In *Proceedings of ITI 2009*, SRCE, Zagreb, 2009, pp. 533–538.

Nivre J, Nilsson J. (2005). Pseudo-Projective Dependency Parsing. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, ACL, 2005, pp. 99–106.

Nivre J. (2006). *Inductive Dependency Parsing*. Springer, 2006.

Nivre J, Hall J, Kübler S, McDonald R, Nilsson J, Riedel S, Yuret D. (2007). The CoNLL 2007 Shared Task on Dependency Parsing. In *Proceedings of the CoNLL Shared Task of EMNLP-CoNLL 2007*, Prague, Czech Republic, pp. 915–932.

Nivre J, Hall J, Nilsson J, Chanev A, Eryigit G, Küble S, Marinov S, Marsi E. (2007). MaltParser: A Language-Independent System for Data-Driven Dependency Parsing. *Natural Language Engineering*, 13(2), 2007, pp. 95–135.

Nivre J, McDonald R. (2008). Integrating Graph-Based and Transition-Based Dependency Parsers. In *Proceedings of ACL 2008: HLT*, ACL, 2008, pp. 950–958.

Pajas P. (2000). Tree Editor TrEd, Prague Dependency Treebank, Charles University, Prague. See URL http://ufal.mff.cuni.cz/tred/ (accessed 2012-10-27).

Passarotti M, Dell'Orletta F. (2010). Improvements in Parsing the Index Thomisticus Treebank: Revision, Combination and a Feature Model for Medieval Latin. In *Proceedings of LREC 2010*, ELRA, 2010, pp. 1964–1971.

Piperidis S. (2012). The META-SHARE Language Resources Sharing Infrastructure: Principles, Challenges, Solutions. In *Proceedings of LREC 2012*, ELRA, 2012, pp. 36–42. See URL http://www.meta-share.eu (accessed 2012-10-27).

Sagae K, Lavie A. (2006). Parser Combination by Reparsing. In *Proceedings of HLT/NAACL*, 2006.

Sgall P, Hajičová E, Panevová J. (1986). *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. Dordrecht, D. Reidel Publishing Company, 1986.

Tadić M. (2002). Building the Croatian National Corpus. In *Proceedings LREC 2002*, ELRA, pp. 441–446.

Tadić M, Fulgosi S. (2003). Building the Croatian Morphological Lexicon. In *Proceedings of the EACL2003 Workshop on Morphological Processing of Slavic Languages*, ACL, 2003, pp. 41–46. See URL http://hml.ffzg.hr (accessed 2012-10-27).

Tadić M. (2007). Building the Croatian Dependency Treebank: The Initial Stages. *Suvremena lingvistika*, 63, pp. 85–92.

Tadić M. (2009). New Version of the Croatian National Corpus. *After Half a Century of Slavonic Natural Language Processing*, Masaryk University, Brno, 2009, pp. 199–205. See URL http://hnk.ffzg.hr (accessed 2012-10-27).

Zeman D. (2002). Can Subcategorization Help a Statistical Dependency Parser? In *Proceedings of COLING 2002*, volume 1, pp. 1–7.