Exploring Pattern Structures of Syntactic Trees for Relation Extraction

Artuur Leeuwenberg*, Aleksey Buzmakov, Yannick Toussaint, and Amedeo Napoli

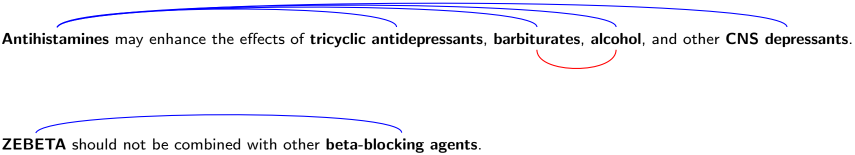
Équipe Orpailleur
LORIA (CNRS – INRIA Nancy Grand Est – Université de Lorraine)

Email: t.leeuwenberg@gmail.com*, firstname.lastname@loria.fr

ICFCA'15, Nerja, June 23-26

Task: Drug-Drug Interaction (DDI) Extraction

Given a sentence, where the drugs are (automatically) annotated, find the pairs of drugs that interact.



Antihistamines may enhance the effects of **tricyclic antidepressants**, **barbiturates**, **alcohol**, and other **CNS depressants**.

ZEBETA should not be combined with other **beta-blocking agents**.

Methods in DDI Extraction

Ensemble Learning (P. Thomas et al. 2011)

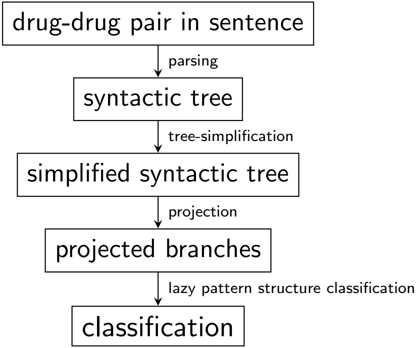
Combination of different kernel based systems (majority voting):

- all-paths graphs (dependency trees)
- k-band shortest path (dependency trees)
- shallow linguistic features

Cased based Reasoning (Maora):

- Each pair is represented by a context and a set of features (lemma, POS, roles)

Pipeline of our approach


drug-drug pair in sentence

parsing

syntactic tree

tree-simplification

simplified syntactic tree

projection

projected branches

lazy pattern structure classification

classification

The DDI Dataset

The dataset used, is of the DDI extraction challenge 2011.[1]

- Build from Drugbank articles
- Drugs are tagged automatically
- Interactions are extracted from the DrugBank, and were manually checked by two domain experts
- Around 4.000 sentences, containing around 2.300 positive and 20.000 negative interactions.

[1] http://labda.inf.uc3m.es/DDIExtraction2011/dataset.html

Entity blinding

- Drugs of the drug-drug pair are replaced with ‘`drug_tag_r`’
- Other drugs are replaced with ‘`drug_tag`’

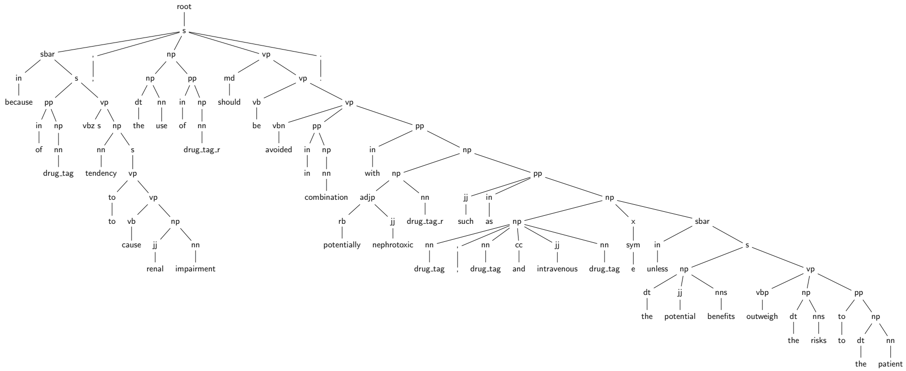
Because of **foscarnets** *tendency to cause renal impairment, the use of* **FOS-CAVIR** *should be avoided in combination with potentially* **nephrotoxic drugs** *such as* **aminoglycosides**, **amphotericin B** *and* **intravenous pentamidine** *unless the potential benefits outweigh the risks to the patient.*

Because of **drug_tag** *tendency to cause renal impairment, the use of* **drug_tag_r** *should be avoided in combination with potentially* **drug_tag_r** *such as* **drug_tag**, **drug_tag** *and* **drug_tag** *unless the potential benefits outweigh the risks to the patient.*

Then the syntax tree is constructed (Stanford Constituency Parser)

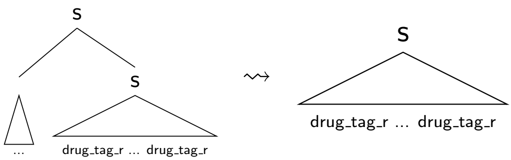
Size of Syntactic Trees

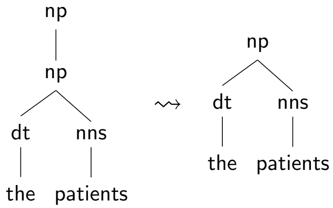
Contains 122 nodes (130 on average, 311 max)


root
s
sbar , np vp .
in s , np pp md vp .
because pp vp dt nn in np should vb vp
in np vbz s np the use of nn be vbn pp pp
of nn nn s drug_tag_r avoided in np in np
drug_tag tendency vp in nn with np pp
to vp combination adjp nn jj in np
to vb np rb jj drug_tag_r such as np x sbar
cause jj nn potentially nephrotoxic nn , nn cc jj nn sym in s
renal impairment drug_tag , drug_tag and intravenous drug_tag e unless np vp
dt jj nns vbp np pp
the potential benefits outweigh dt nns to np
the risks to dt nn
the patient

Tree Simplification

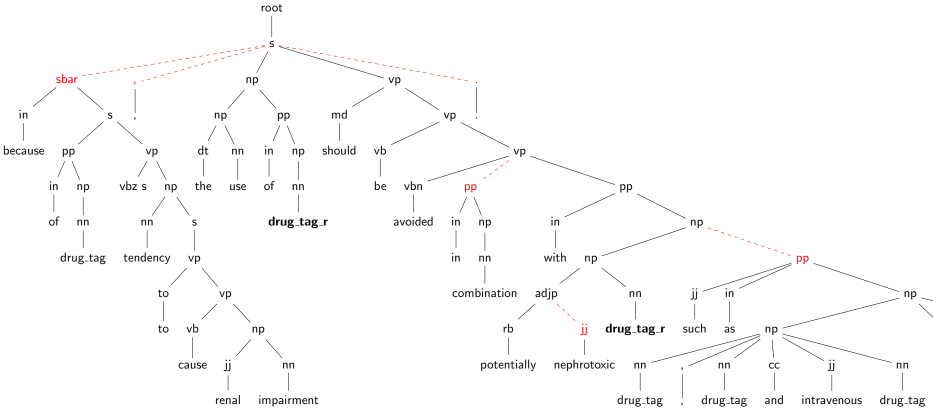
1. Removal of JJ, PP, S, SBAR, PRN subtrees that don't contain 'drug_tag_r'
2. VP-nodes become NEGVP if they contain a negation.
3. Only the lowest S node containing both 'drug_tag_r' tags is considered.
4. Contraction of non-branching trees





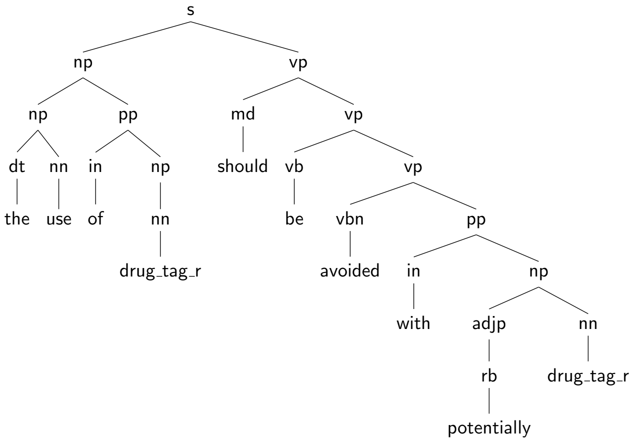
[1] JJ=adjective, PP=prepositional phrase, S/SBAR=sentence, PRN=bracketed expression, VP=verb phrase

Simplifications



Simplified Tree

Contains 31 nodes (41 on average, 138 max)



Pattern Structures (B. Ganter and S. Kuznetsov 2001)

A *pattern structure* is defined as a tuple $(G, (D, \sqcap), \delta)$

- G is the set of objects
- D is the set of object descriptions (or patterns)
- $\delta : G \to D$ maps objects to their corresponding description
- $\sqcap$ is a similarity operator on subsets of D (idempotent, associative, and commutative)

The subsumption relation between subsets of descriptions can be defined in a standard way:

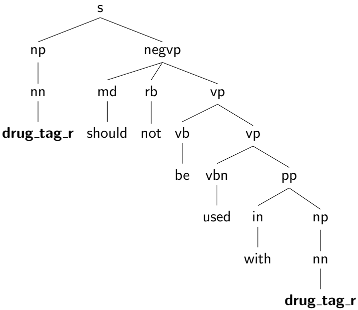
$$A \sqsubseteq B :\Longleftrightarrow A \sqcap B = A$$

Pattern Structure of Syntactic Trees $(G, (D, \sqcap), \delta)$

Object

Anafranil should not be used with **MAO inhibitors**. (d16.s2.p0)

Object Description

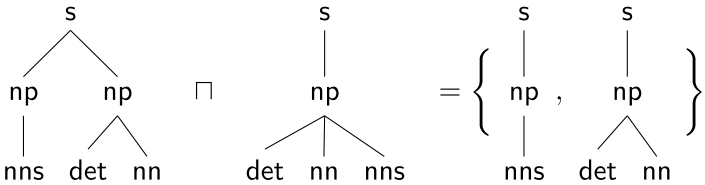


Pattern Structure of Syntactic Trees $(G, (D, \sqcap), \delta)$

Rooted Tree Intersection ($\sqcap$)

The *Rooted Intersection* between tree t_1 and tree t_2 is the set of maximal trees from the intersection of all rooted subtrees of t_1 and those of t_2.

Example



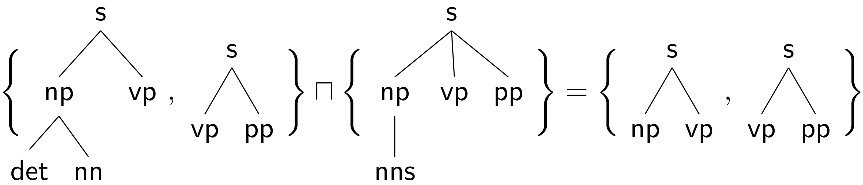
Pattern Structure of Syntactic Trees $(G, (D, \sqcap), \delta)$

Similarity Operator ($\sqcap$)

The *similarity* between a set of trees A and a set of trees B ($A \sqcap B$) is the set of maximal trees from

$$\bigcup_{(t_a, t_b) \in A \times B} t_a \sqcap t_b$$

Example

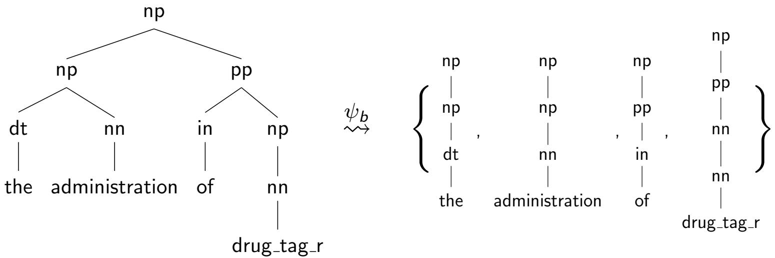


Projection for the Pattern Structure of Syntactic Trees

Branch Projection

The branch projection of a tree t is the set of its branches $\psi_b(t)$.

Example

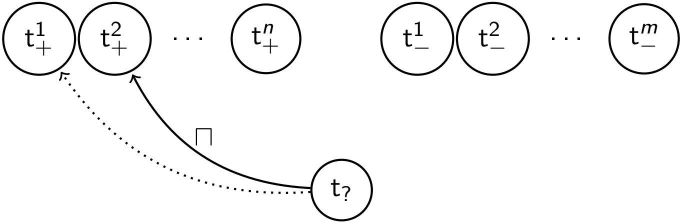


Lazy Pattern Structure Classification (Kuznetsov 2013)

Given:

- set of positive examples T_+
- set of negative examples T_-

A new object $t_?$ is classified as positive iff a positive hypothesis can be found in T_+.

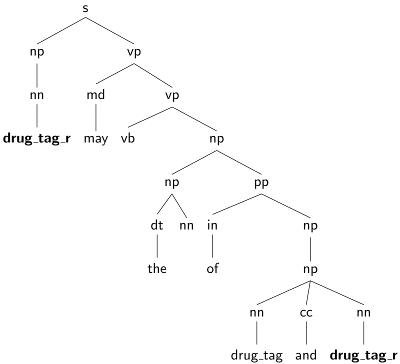


$$\exists_{t_+ \in T_+} \forall_{t_- \in T_-} : (\delta(t_?) \sqcap \delta(t_+)) \not\sqsubseteq \delta(t_-)$$

Example of a Positive Hypothesis

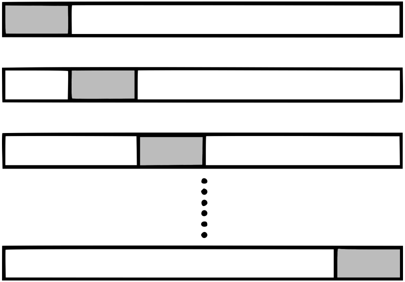
$\underline{t_?}$: **Ketoconazole** tablets may alter the metabolism of cyclosporine, tacrolimus, and **methylprednisolone**, resulting in elevated plasma concentrations of the latter drugs.

$\underline{t_+}$: **Ethoxzolamide** may increase the action of tricyclics, amphetamines, **procainamide**, and quinidine.



10-fold Cross Validation

We did a 10-fold cross validation on the 23.000 drug-drug pairs.



We calculate precision, recall and f-measure over all the folds.

Results

Simplifications	P	R	F_1
1. NEGVP, contraction	0.32	0.39	0.35
2. lowest-S, contraction	0.27	**0.49**	0.35
3. NEGVP, lowest-S	0.36	0.45	0.38
4. NEGVP, lowest-S, contraction	0.30	0.49	0.37
5. NEGVP, lowest-S, vp-map	0.35	0.45	0.39
6. NEGVP, lowest-S, vp-map, prep-map	**0.39**	0.41	**0.40**

Table: Results from 10-fold cross validation on the DDI 2011 data set. Performance is measured in precision (P), recall (R) and F_1-measure (F_1). In all conditions constituent simplification is applied.

Error Analysis

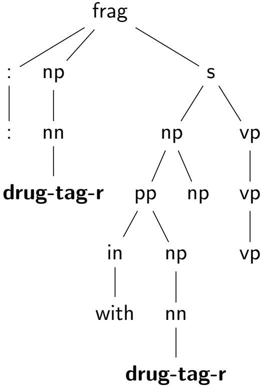
Error categories that we found:

- Non-sentences (NP, FRAG)
- Very deep syntactic trees (mostly FN)
- Mistakes in annotation (mostly FP)
- Insufficient similarity
- Parsing errors

False Positive

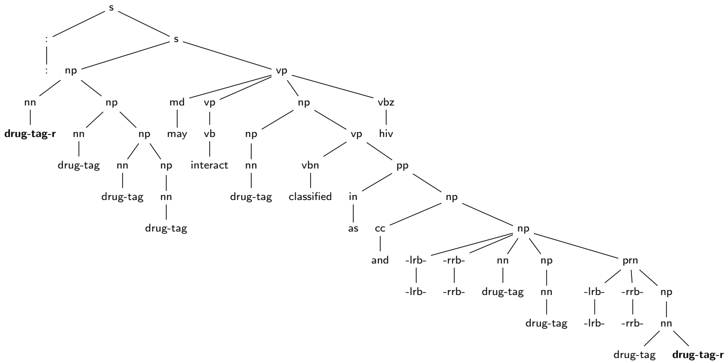
$\underline{t_?}$: **Thalidomide**: Co-administration with **thalidomide** should be employed cautiously, as toxic epidermal necrolysis has been reported with concomitant use.

$\underline{t_+}$: **Corticosteroids**: Concomitant administration with **aspirin** may increase the risk of gastrointestinal ulceration and may reduce serum salicylate levels



False Negative

$t_?$: **Etonogestrel** may interact with the following medications: acetaminophen (Tylenol), antibiotics such as ampicillin and tetracycline, anticonvulsants (Dilantin, Phenobarbital, Tegretol, Trileptal, Topamax, Felbatol), antifungals (Gris-PEG, Nizoral, Sporanox), atorvastatin (Lipitor), clofibrate (Atromid-S), cyclosporine (Neoral, Sandimmune), HIV drugs classified as protease inhibitors (Agenerase, Crixivan, Fortovase, Invirase, Kaletra, Norvir, Viracept), morphine (Astramorph, Kadian, **MS Contin**), phenylbutazone, prednisolone (Prelone), rifadin (rifampin), St. Johns wort, temazepam, theophylline (Theo-Dur), and vitamin C.



Conclusions & Future Work

Proposed:

- Pattern structure of syntactic trees
- Branch-projection
- Extraction of characteristic syntactic tree patterns for classification

Future work:

- Filter badly performing patterns
- Use the patterns in different classification paradigms
- Improve the tree simplifications
- Include morphological or semantic information in the trees
- Application to dependency graphs, or parse thickets

Exploring Pattern Structures of Syntactic Trees for Relation Extraction

Artuur Leeuwenberg*, Aleksey Buzmakov, Yannick Toussaint, and Amedeo Napoli

Équipe Orpailleur
LORIA (CNRS – INRIA Nancy Grand Est – Université de Lorraine)

Email: t.leeuwenberg@gmail.com*, firstname.lastname@loria.fr

ICFCA'15, Nerja, Ju

References I

Balcázar, J. L., Bifet, A., & Lozano, A. (2006). Intersection algorithms and a closure operator on unordered trees. *MLG*, 1.

Chowdhury, F. M., Abacha, A. B., Lavelli, A., & Zweigenbaum, P. (2011). Two different machine learning techniques for drug-drug interaction extraction. *Challenge Task on Drug-Drug Interaction Extraction*, 19–26.

Chowdhury, M. F. M., & Lavelli, A. (2011). Drug-drug interaction extraction using composite kernels. *Challenge Task on Drug-Drug Interaction Extraction*, 27–33.

Ganter, B., & Kuznetsov, S. O. (2001). Pattern structures and their projections. In *Conceptual structures: Broadening the base* (pp. 129–142). Springer.

Kuznetsov, S. O. (2013). Fitting pattern structures to knowledge discovery in big data. In *Formal concept analysis* (pp. 254–266). Springer.

References II

Kuznetsov, S. O., & Samokhin, M. V. (2005). Learning closed sets of labeled graphs for chemical applications. In *Inductive logic programming* (pp. 190–208). Springer.

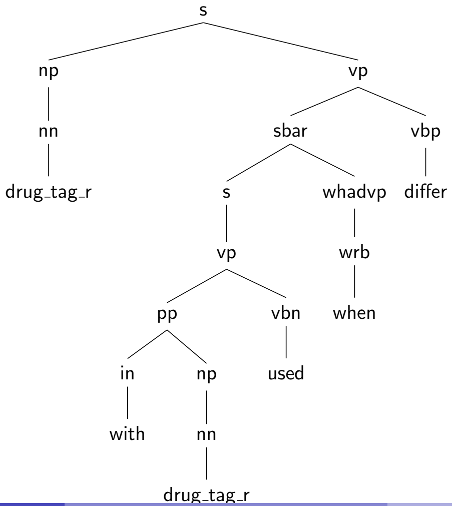
Socher, R., Bauer, J., Manning, C. D., & Ng, A. Y. (2013). Parsing with compositional vector grammars. In *In proceedings of the acl conference.*

Thomas, P., Neves, M., Solt, I., Tikk, D., & Leser, U. (2011). Relation extraction for drug-drug interactions using ensemble learning. *Challenge Task on Drug-Drug Interaction Extraction*, 11–18.

Wille, R. (2009). *Restructuring lattice theory: an approach based on hierarchies of concepts*. Springer.

False Positive

test: It is not known if [**hormonal contraceptives** differ in their effectiveness when used with **Accutane**.]


s
np
vp
nn
sbar
vbp
drug_tag_r
s
whadvp
differ
vp
wrb
pp
vbn
when
in
np
used
with
nn
drug_tag_r