TectoMT

Roman Lacko
Josef Plch

Framework Treex
(dříve TectoMT)

Motivace

- většina aplikací NLP vzniká propojením existujících nástrojů

```
cat $FILE | $TOKENIZER | $LEXER | $PARSER | …
```

- propojení většinou není snadné:
  - nekompatibilní formáty nástrojů ⇒ *obalovací program (wrapper)*
  - rozsáhlé nebo chybějící dokumentace
  - některé nástroje nelze integrovat vůbec (např. vyžadují nekompatibilní platformy)

- často je řešení pouze *jednorázové*

Motivace

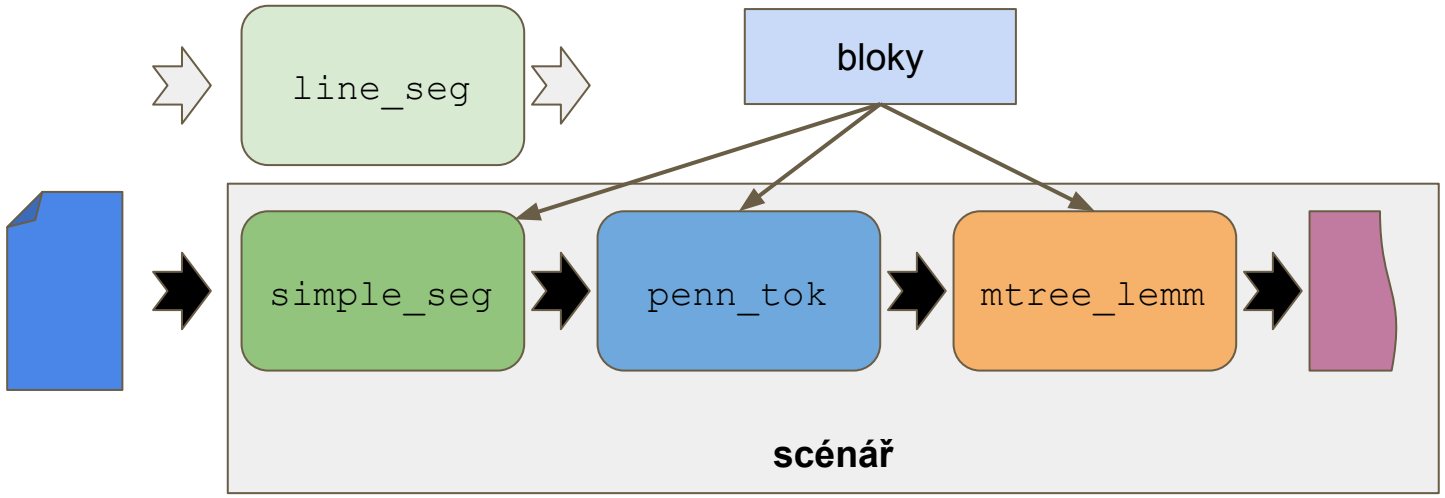
Framework v NLP má za úkol sjednotit používání jazykových nástrojů:

- zajišťuje požadovaný formát dat pro použité nástroje
- definuje API pro snadné přidání dalších nástrojů
- umožňuje se soustředit na zpracování dat

⇒ framework Treex (původně pod názvem **TectoMT**)

Technologie

- důraz na modularitu a *znovupoužitelnost*
- **bloky** ⇒ posloupnost kroků v jednotlivých fázích analýzy
- **scénář** ⇒ konkrétní spojení některých bloků



Technologie

- framework je implementován v jazyce Perl

*„**Suicides** protect your environment.”*

*„Debugging is just a **regular nightmare**, not worse.”*

citováno z [**4**]

Bloky

V roce 2010 obsahoval přes **400 bloků:**

- 140 pro angličtinu
- 120 pro češtinu
- 60 pro překlad angličtina ⇒ čeština
- 30 pro jiné jazyky
- 50 jazykově nezávislých

Bloky zahrnují například tyto nástroje:

- **angličtina**: 5 značkovačů, 2 složkové a 2 závislostní parsery, 1 NER
- **čeština**: 3 značkovače, 3 závislostní parsery, 2 NER
- **němčina**: 1 značkovač, 2 závislostní parsery

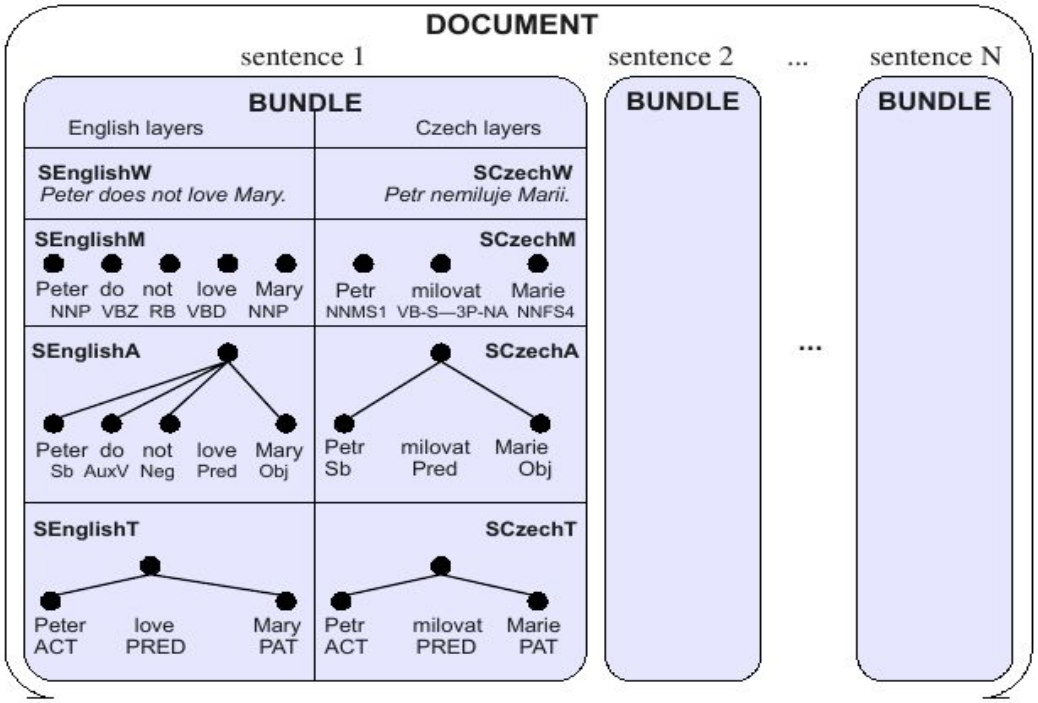
Aplikace v TectoMT

Zpracování textu většinou sestává ze stejné posloupnosti kroků:

1. **překlad vstupních dat** do standardního tvaru pro Treex
2. **aplikace scénaře** na data
3. **překlad výsledku** ve formátu Treex do požadovaného tvaru

Dokumenty, svazky a stromy

- dokument (paralelní sekvence vět) je uložen v jednom souboru
- každá věta je reprezentována svazkem (*bundle*) stromů



Překladový systém TectoMT

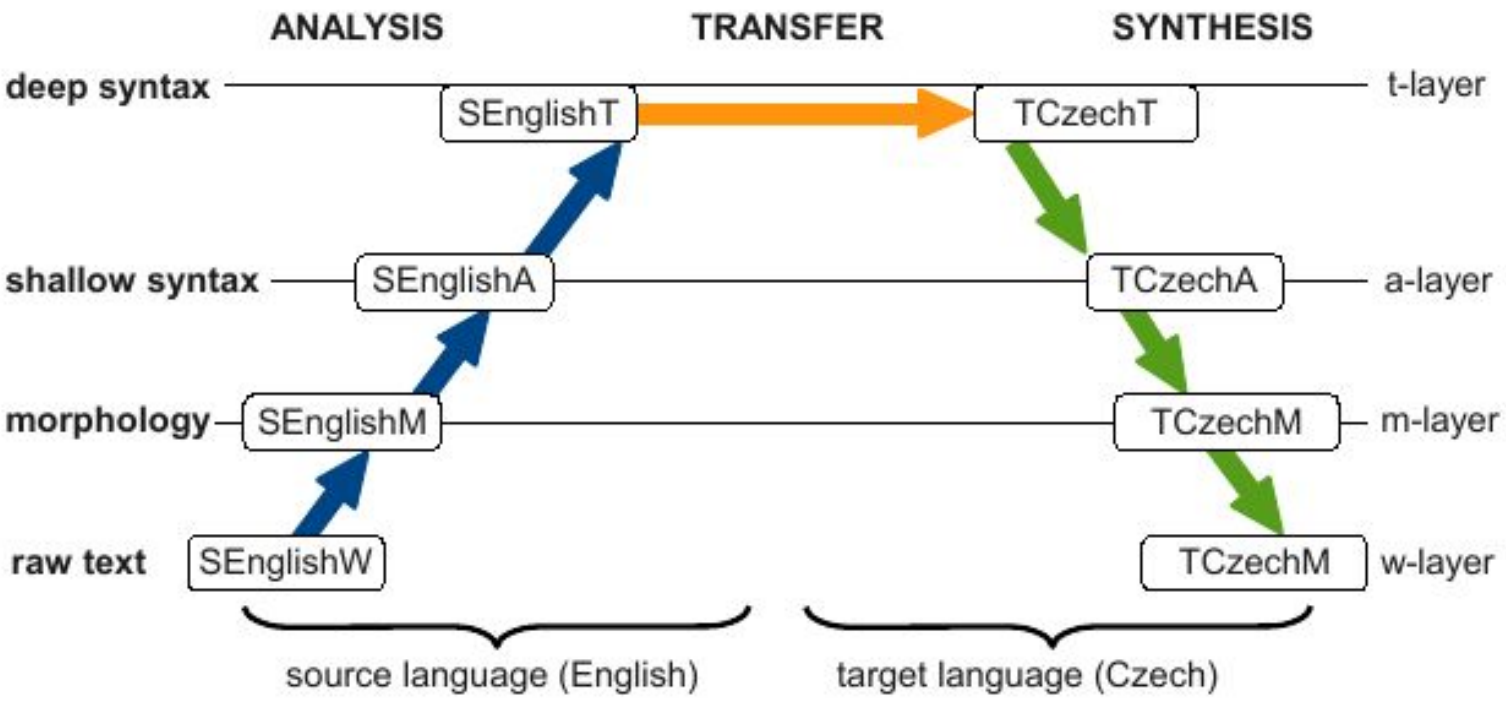
Co je TectoMT

- systém strojového překladu kombinující pravidlové i statistické prvky
- založen na architektuře

analýza ⇒ transfer ⇒ syntéza

- transferní vrstvu tvoří *tektogramatická rovina*

Úrovně popisu jazyka



Úrovně popisu jazyka

Převzato z PDT (*Prague Dependency Treebank*) 2.0

1.	**word layer**	čistý text
2.	**morphological layer**	sekvence označkovaných pozic
3.	**analytical layer**	mělké syntaktické stromy
4.	**tectogrammatical layer**	hluboké syntaktické stromy

Jednotlivé vrstvy se označují *w-layer*, *m-layer* atd.

Kromě toho má TectoMT **phrase-structure layer** na stejné úrovni jako *a-layer*.

Tektogramatická rovina

Stavebními kameny jsou **t-stromy** (*hluboké závislostní stromy*)

- klasická gramatická rovina zachycuje vztahy mezi všemi slovy ve větě
- uzly v tektogramatické rovině (hloubkové syntaxi) tvoří pouze významová slova
- informace z pomocných uzlů (například pomocná slovesa) tvoří **formémy** (atributy) nadřazených **t-uzlů**
- hrany t-stromu představují lingvistické závislosti

Příklady formémů:

- `n:subj` podstatné jméno v pozici předmětu
- `n:for+X` podstatné jméno s předložkou *for*
- `v:because+fin` sloveso v podřazené věte s *because*

Příklad | Analýza

`w-english` ⇒ `m-english`

- rozdělení vět na pozice (tokeny)
- označkování a lematizace

`m-english` ⇒ `a-english`

- označení hlavního uzlu (*head node*) pomocí heuristik
- konstrukce klasického závislostního stromu

Příklad | Analýza

`a-english` ⇒ `t-english`

- nalezení pomocných a-uzlů (předložky, spojky, …)
- sestavení t-stromů
- nalezení kořenových t-uzlů v t-podstromech
- dopočtení pomocných atributů (klasifikace uzlů, gramatémy,...)

Příklad | Transfer a syntéza

`t-english` ⇒ `t-czech`

- překlad lemmat t-uzlů (použije statisticky nejčastější protějšek)
- zjištění osoby a vidu
- nalezení koreferencí mezi zájmeny a předměty

`t-czech` ⇒ `a-czech`

- vytvoření dalších a-uzlů
- propagace hodnot atributů
- zajištění shody

`a-czech` ⇒ `w-czech`

- vytvoření vět projekcí a-stromů na přímku

Výsledky

- samostatné TectoMT dosahuje ve srovnání se statistickými překladači výrazně horších výsledků
- poradí si ale s některými problémy, které dělají statistickým nástrojům potíže
- kombinace TectoMT a statistického překladače Moses (pod názvem Chimera) dosahuje lepších výsledků než samotný Moses
- Chimera v letech 2013–2015 zvítězila v soutěži překladačů „Workshop on Statistical Machine Translation“

Reference

1. Žabokrtský, Zdeněk, Jan Ptáček, and Petr Pajas. "*TectoMT: Highly modular MT system with tectogrammatics used as transfer layer.*" Proceedings of the Third Workshop on Statistical Machine Translation. Association for Computational Linguistics, 2008.
2. Popel, Martin, Zdeněk Žabokrtský. "*TectoMT: Modular NLP Framework.*" Advances in Natural Language Processing, 7th International Conference on NLP, IceTAL 2010.
3. Popel, Martin. "*TectoMT: Deep-Syntactic Machine Translation*" [Internet]. 2015 (navštíveno 8. 11. 2015). Odkaz na prezentaci.
4. Bojar, Ondřej. "*TectoMT for Plaintext Freaks*" [Internet]. 2009 (navštíveno 8. 11. 2015). Odkaz na prezentaci.