Automatic Alignment of Tectogrammatical Trees from Czech-English Parallel Corpus

David Mareček

obhajoba diplomové práce

8. 9. 2008

Motivace

- Na t-rovině jsou si jazyky podobnější $\Rightarrow$ alignment by zde měl být „jednodušší“ než na m-rovnině.
- Zlepšení transferu pro systém automatického překladu TectoMT
- Vyhledávání překladových ekvivalentů mezi dvěma jazyky

Cíl práce

- Implementovat nástroj, který dostane na vstupu českou a anglickou větu analyzovanou až na t-rovinu a vrátí seznam navzájem si odpovídajících uzlů jejich t-stomů
- Měl by využívat stromových struktur, ale zároveň by měl být dostatečně odolný proti chybám vzniklým během parsingu
- Implementace v prostředí TectoMT

Zjednodušená t-rovina

- Strutura t-roviny převzatá z TectoMT
  - Chybí aktuální členění (pořadí uzlů v t-stromu odpovídá pořadí slov ve větě)
  - Chybí „kopírované uzly“
  - Chybí odkazy do valenčního slovníku
  - Chybí funktory (kromě koordinačních)

Vytvoření testovacích dat

- ruční párování na slovní rovině
- následný převod na t-rovinu pomocí odkazů **`lex.rf`**

Manuální word-alignment

- K dispozici 515 ručně anotovaných vět z PCEDT
- Tři typy hran: *sure, phrasal, possible*
- Anotace dalších 1985 vět z corpusu CzEng
- 4 anotátoři (MFF, FF, 2x ÚJČ)
- Každá věta anotována dvěma anotátory

typ dat	vět	EN tokenů	CZ tokenů
Právní texty (Acquis Communautaire)	501	13 512	10 752
Beletrie (Reader's Digest, E-Books)	500	10 097	8 978
Novinové články (Project Syndicate)	484	10 714	9 990
Nov. Čl. s pojmenovanými entitami (Project Syndicate)	500	12 799	11 052

Příklady word-alignmentu

Převod do t-roviny

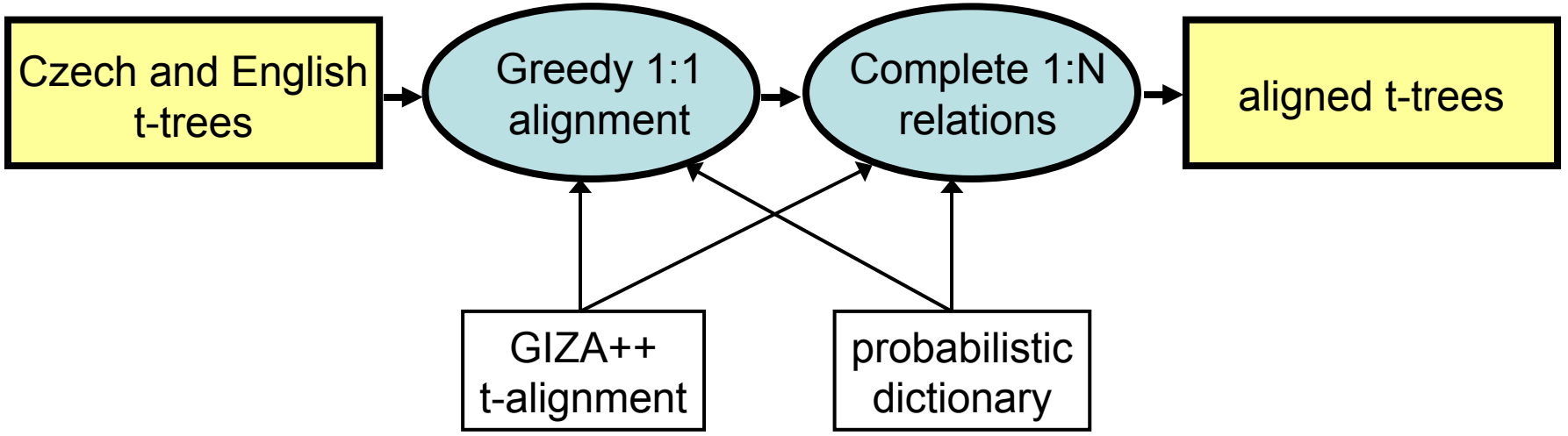
- využití atributu **`lex.rf`**
- dva uzly na t-rovině jsou spojeny hranou ⇔ jejich odpovídající tokeny na m-rovině jsou spojeny hranou
- typy hran jsou zachovány
- problém: přidané uzly, např. #PersPron reprezentující nevyjádřený podmět

Mezianotátorská shoda

typ dat	typy hran rozlišovány		typy hran nerozlišovány		pouze hrany „sure“	
rovina	m	t	m	t	m	t
právní texty	86,7	**92,0**	92,1	**96,1**	94,0	**95,6**
beletrie	76,6	**82,9**	85,8	**90,5**	90,1	**91,2**
novinové články	80,8	**87,9**	87,8	**93,3**	92,0	**93,7**
pojmenované entity	86,5	**94,3**	91,5	**96,5**	93,6	**96,4**
PCEDT	84,1	**94,3**	89,8	**95,1**	93,8	**95,3**

T-aligner

- Implementován v prostředí TectoMT
- 2 fáze



Fáze 1 – Greedy 1:1 alignment

- hladový algoritmus založený na rysech (features)
- skóre potenciálního páru t-uzlů:

$$\mathbf{score(cnode, enode) = \vec{w} \cdot \vec{f}(cnode, enode)}$$

váha hodnota

- algoritmus vybere vždy ten pár s nejvyšším skóre a příslušné uzly spojí
- to opakuje, dokud je skóre větší než daná prahová hodnota

Rysy (features)

- vlastnosti dvojice českého a anglického t-uzlu
- celkem 15 rysů
- jejich váhy byly optimalizovány pomocí perceptronu
- příklady rysů:
  - Pravděpodobnost překladu dvojice t-lemmat (z pravděpodobnostního slovníku) (hodnota 0 až 1)
  - Tato dvojice rovněž byla/nebyla spárována nástrojem GIZA++ (0 nebo 1)
  - T-lemmata jsou shodná (0 nebo 1)
  - Shoda t-lemmat v prvních třech písmenech (0, 1)

Rysy (features) (2)

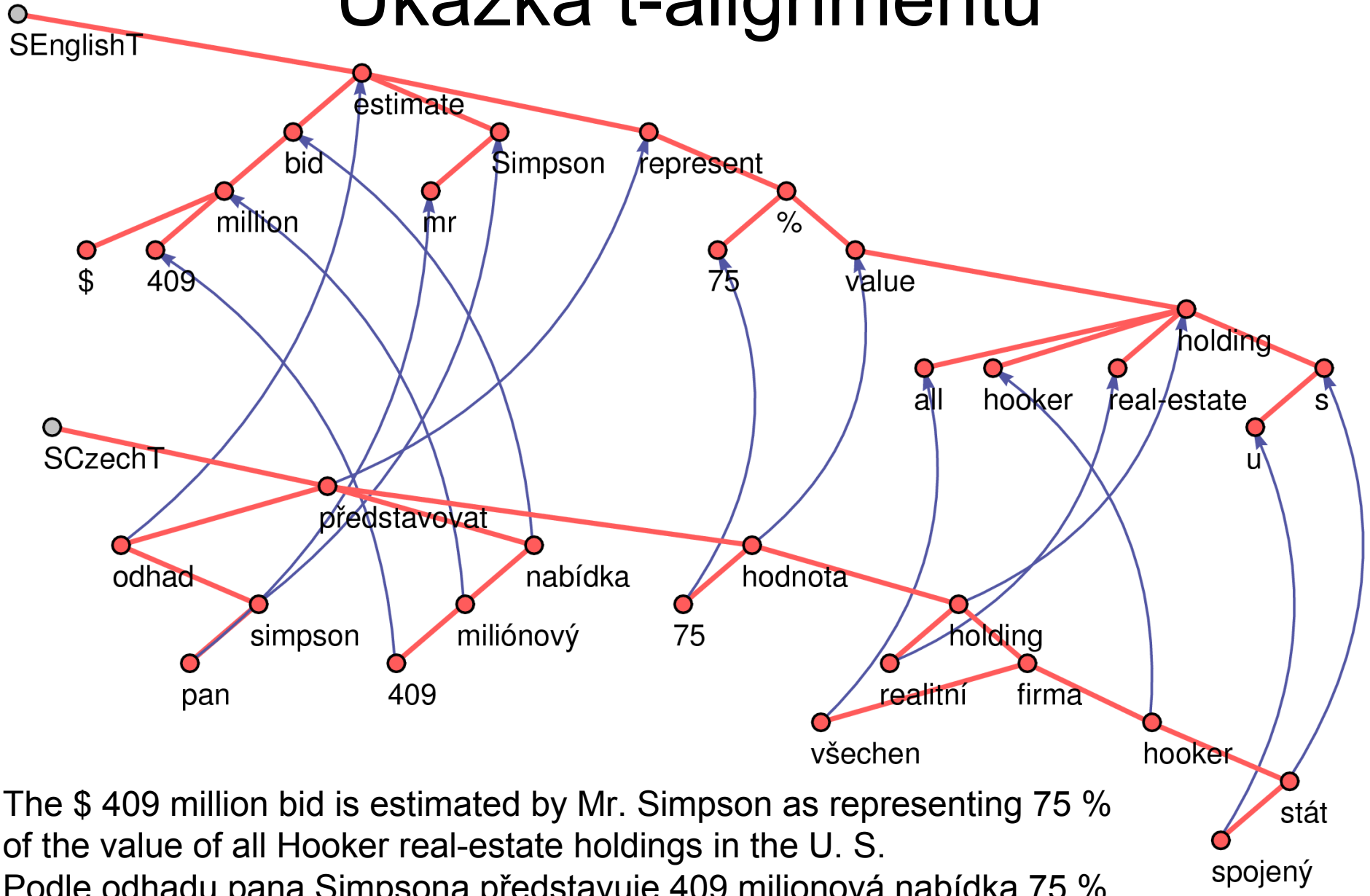
- Rodičovské uzly obou zkoumaných uzlů jsou již spárovány (hodnoty 0, 1)
- Někteří potomci zkoumaných uzlů jsou již spárováni (hodnoty 0, 1, 2, 3 …)
- Shodný sémantický slovní druh (hodnoty 0, 1)
- Podobná pozice uzlů ve stomě (real 0 až 1)
  - Rozdíl relativních pozic českého a anglického uzlu

Fáze 2 – Completing 1:N relations

Spojíme navíc ty dvojice uzlů (K, L), pro které jsou splněny následující podmínky:

- K ještě není spojen s žádným protějšímu uzlem, ale jeho synovský nebo rodičovský uzel je spojen s L.
- Dvojice (K, L) byla navíc spojena systémem GIZA++
- Dvojice (K, L) se vyskytuje v pravděpodobnostním slovníku

Ukázka t-alignmentu



The $ 409 million bid is estimated by Mr. Simpson as representing 75 % of the value of all Hooker real-estate holdings in the U. S.
Podle odhadu pana Simpsona představuje 409 milionová nabídka 75 % hodnoty všech realitních holdingů firmy Hooker ve Spojených státech.

Vyhodnocení úspěšnosti

- pro porovnání výsledků využita pouze spojení typu „sure“.
- 10-násobná cross-validace
- 9/10 dat pro natrénování optimálních vah jednotlivých rysů (features) perceptronem
- 1/10 dat pro samotnou evaluaci

T-aligner – evaluace

typ dat	precision	recall	f-measure
právní texty	92,8 %	91,4 %	92,1 %
beletrie	86,3 %	82,7 %	84,4 %
novinové články	90,7 %	91,9 %	91,3 %
pojmenované entity	94,8 %	92,8 %	93,8 %
PCEDT	94,9 %	88,7 %	91,7 %
Dohromady	**92,5 %**	**89,6 %**	**91,0 %**

GIZA++ na m-lemmatech

- GIZA++ byla spuštěna na sekvence m-lemmat.
- Průniková symetrizace
- Výsledný word-alignment byl pak převeden na t-alignment pomocí odkazů **`lex.rf`**.

precision	recall	f-measure
95,5 %	77,8 %	85,7 %

GIZA++ na t-lemmatech

- Z tektogramatických stromů byla vyextrahována t-lemmata a seřazena podle jejich pozice ve stromě.
- Každý uzel reprezentuje jedno t-lemma.
- Na takovéto sekvence t-lemmat byla spuštěna GIZA++ s průnikovou symetrizací.

precision	recall	f-measure
93,1 %	75,9 %	83,6 %

Závěr

- zlepšení f-measure o 5% oproti baseline (GIZA++)

	f-measure
GIZA++ t-alignment	85,7 %
feature based t-aligner	**91,0 %**
mezianotátorská shoda	94,8 %

Děkuji za pozornost.