Porovnání úspěšnosti tagování korpusu[1]

Hana Skoumalová
Ústav teoretické a komputační lingvistiky, Filozofická fakulta
Univerzita Karlova

Abstract

In this paper, we present the way how Czech National Corpus is morphological tagged and then we evalute the accuracy of the tagging. Because there is no testing corpus with the same tagset, and also because we want to compare several methods of tagging, we developed a method for comparing corpora with slightly different tagsets. In the end, the accuracy of all presented tagging methods is measured and future improvment of tagging is proposed.

1. Úvod

Při automatickém tagování (morfologickém značkování) korpusu je důležité vědět, jak je tento proces úspěšný, neboli jaké je procento chyb v otagovaném textu. Provést úplnou kontrolu tagování by prakticky znamenalo totéž jako otagovat celý korpus ručně, což je, zejména u velkých korpusů, úkol nad lidské síly. Prakticky se proto postupuje tak, že se ručně otaguje pouze menší část (několik desítek tisíc slov), tento testovací soubor se nechá otagovat i automaticky a výsledky ručního a automatického tagování se porovnají. Ručně tagovaný korpus se přitom považuje za dokonalý, i když i v něm se můžou vyskytovat chyby. Úspěšnost tagování naměřená na tomto testovacím korpusu se pak považuje za úspěšnost měřené metody.

Pro tagování Českého národního korpusu (ČNK)[2] se využívá kombinace několika metod, kterými se postupně desambiguuje výstup z morfologické analýzy. Poslední evaluace tagování byla provedena v roce 2007 (Spoustová et al. 2007), avšak od té doby se poněkud změnily některé metody i použitý tagset. V tomto článku tedy popíšeme metody, kterými se značkuje ČNK v současnosti, způsob jejich evaluace a porovnání s dříve naměřenými výsledky.

2. Tagování korpusu

Pro tagování ČNK se používá posloupnost několika procedur, které každému slovnímu tvaru, číslu a interpunkčnímu znaménku přiřadí značku popisující slovní druh a morfologické charakteristiky. Existuje několik možností, jak tohoto cíle dosáhnout. Jednou je stochastická metoda, která přiřazuje jednotlivým slovům tagy (značky) s morfologickými charakteristikami. Tato metoda nepoužívá žádný slovník, ale pouze trénovací korpus, ve kterém jsou značky přiřazeny manuálně.

Další možností je využít morfologický slovník, s jehož pomocí se nejdříve všem tvarům přiřadí lemmata a tagy odpovídající morfologické interpretace, a poté se z těchto

[1] Práce na tomto příspěvku byla podpořena grantem GAČR, reg. č. P406/10/0434 a výzkumným záměrem MSM 0021620823

[2] viz http://www.korpus.cz

lemmat a tagů vybírá jedna dvojice, která odpovídá použití tvaru v dané větě. Tento proces se nazývá desambiguace a je zde opět několik možností jak desambiguaci provést: stochastickými metodami, lingvistickými pravidly, nebo kombinací těchto metod.

Pro značkování ČNK byla zvolena metoda s využitím morfologického slovníku a s následnou desambiguací hybridní metodou. Jednotlivé kroky budou teď popsány podrobněji.

2.1. Morfologická analýza

První procedurou je morfologická analýza slovních tvarů (Hajič 2004). Z praktických důvodů tato procedura zahrnuje i tzv. tokenizaci textu, což znamená rozdělení textu na věty, jednotlivé slovní tvary a interpunkci. Výsledkem morfologické analýzy je tedy text rozdělený na jednotlivá slova, ke kterým jsou přiřazena lemmata a tagy. Tagy mají konstantní délku a každá z 15 pozic má daný význam (viz Tabulka 1).

	Název pozice	Význam
1	POS	slovní druh
2	SUBPOS	poddruh
3	GENDER	jmenný rod
4	NUMBER	číslo
5	CASE	pád
6	POSSGENDER	rod posesora
7	POSSNUMBER	číslo posesora
8	PERSON	osoba
9	TENSE	čas
10	GRADE	stupeň
11	NEGATION	negace
12	VOICE	slovesný rod
13	RESERVE1	
14	RESERVE2	
15	VAR	tvarová varianta

Tabulka 1: Pozice v morfologickém tagu

2.2. Bezpečná pravidla

Následuje procedura obsahující lingvistická pravidla (Petkevič 2006), (Květoň 2006). Výsledkem použití pravidel je vymazání některých lemmat nebo tagů z množiny hypotéz.

Pravidla jsou formulována tak, aby pokud možno nedošlo k vymazání správné interpretace slovní formy; v případě, že je věta příliš složitá, ponechají se všechny tagy.

2.3. Frazémy

Následuje modul, který v textu vyhledává frazémy a na základě jejich identifikace maže další nevyhovující lemmata a tagy. Vyhledávané frazémy jsou ustálená slovní spojení s pevným nebo volným slovosledem, části frází nebo celé věty. Více o proceduře viz (Hnátková 2002).

2.4. Pravidla s heuristikami

V následném kroku se opět spustí lingvistická pravidla, tentokrát však zahrnují i některé heuristiky, u kterých není stoprocentní jistota, že nevymažou správnou interpretaci; tato možnost je však velmi nepravděpodobná. Řeší se zde homonymie jednotlivých slovních tvarů jako např. *branou* (lemma *brána* x *braný*), *sacích* (*sak* x *sací*), *uhranou* (*uhranout* x *uhraný*) apod.

2.5. Stochastický tagger

I po předchozích procedurách zbývají nerozhodnuté případy, kdy u jednoho slovního tvaru máme několik lemmat nebo tagů. Závěrečnou fázi desambiguace tak provede stochastický tagger Morče (Votrubec 2006), který ke každému tvaru vybere právě jedno lemma a jeden tag.

3. Předchozí evaluace

Při evaluaci provedené v roce 2007 (Spoustová 2007) byly porovnávány hlavně různé stochastické taggery, konkrétně HMM tagger (Krbec 2005), tagger založený na rysech (Hajič et al. 2006) (dále nazývaný JH tagger) a Morče. Byly porovnávány samostatně, ale i v součinnosti s lingvistickými pravidly. V době, kdy bylo porovnání prováděno, neobsahovala lingvistická procedura ještě frazémový modul, a proto byla pravidla spouštěna jen jednou.

Při evaluaci byla měřena jednak úspěšnost jednotlivých stochastických taggerů a jednak úspěšnost různých kombinací taggerů a pravidel. Měřeny byly tyto kombinace metod:

1. lingvistická pravidla a stochastický tagger
2. určení slovního poddruhu libovolným stochastickým taggerem, poté lingvistická pravidla a týž tagger jako na počátku
3. sjednocení výsledků tří stochastických taggerů a posléze lingvistická pravidla a stochastický tagger

Trénování stochastických taggerů i evaluace byly provedeny na Pražském závislostním treebanku (PDT v2.0) (Hajič et al 2006) a byl při tom použit modul morfologické analýzy ve verzi z dubna 2006.

Deklarovaná úspěšnost jednotlivých stochastických taggerů je uvedena v Tabulce 2, úspěšnost kombinovaných metod je uvedena v Tabulce 3.

tagger	úspěšnost
JH tagger	94,04 %
HMM tagger	94,82 %
Morče	95,12 %

Tabulka 2: Úspěšnost jednotlivých stochastických taggerů

kombinace metod	úspěšnost
bezpečná pravidla+Morče	95,34 %
SUBPOS+bezpečná pravidla+Morče	95,44 %
sjednocení 3 taggerů+stochastický tagger	95,52 %
sjednocení 3 taggerů+pravidla včetně heuristik+stochastický tagger	95,68 %

Tabulka 3: Úspěšnost kombinovaných metod naměřená v r. 2007

4. Rozdíly v tagování mezi r. 2007 a 2011

Tagování ČNK se od r. 2007 změnilo v několika ohledech (Jelínek 2008). V prvé řadě se zlepšila morfologická analýza, která se průběžně obohacuje o nová slova a zároveň se v ní opravují chyby objevené při využívání korpusu.

Další změna se týká hodnot na některých pozicích v tagu. Tagset použitý v morfologické analýze obsahuje některé "proměnné" s významem libovolná hodnota, popř. libovolná hodnota z nějaké dané množiny (např. X na třetí pozici vyhrazené pro jmenný rod má význam *libovolný rod*, Y na téže pozici má význam *mužský rod životný nebo neživotný*, naproti tomu X na čtvrté pozici vyhrazené pro číslo má význam *libovolné číslo*). Pravidla používaná v procesu desambiguace se snaží tyto víceznačnosti odstranit a zvolit právě jednu hodnotu z nabízené množiny.

Další změnou oproti klasickému tagsetu je zrušení některých hodnot v určitých tazích, např. číslo u zvratného zájmene/částice nebo nebo čas u příčestí. Tyto hodnoty, které byly obsaženy v původním tagsetu, nevyjadřují morfologické vlastnosti určovaného slovního tvaru a měly sloužit pro snazší strojové zpracování otagovaného textu při následné syntaktické analýze. Protože však Český národní korpus má sloužit přímo uživatelům, rozhodli jsme se tyto hodnoty z tagů vypustit.

Poslední významnou změnou je přehodnocení slovního druhu u některých adverbií na částice (např. *beztak*, *namouduši*, *nesporně* atd.). Toto přehodnocení bylo provedeno pouze pro ČNK a v programu pro morfologickou analýzu zůstaly původní hodnoty. Proto se tyto změny realizují zvláštní procedurou spouštěnou na závěr celého procesu desambiguace.

5. Nová evaluace tagování

Postup tagování popsaný v oddílu 2 nebyl doposud evaluován, ačkoliv se již několik let používá pro značkování ČNK. Proto jsme se rozhodli provést novou evaluaci a porovnat ji s výsledky dosaženými při předchozím měření úspěšnosti.

5.1. Evaluace různých metod tagování

Jak již bylo řečeno výše, nová metoda značkování korpusů se od těch předešlých liší především použitím mírně odlišného tagsetu. Aby bylo možno takto rozdílné metody porovnat, byla vytvořena aplikace, která dokáže porovnat tagy ze dvou různých tagsetů.
Hlavní myšlenkou této aplikace je, že hodnoty vyskytující se na jednotlivých pozicích v tagu představují buď atomickou hodnotu nebo množinu hodnot. Porovnáváme-li tedy jednu pozici ve dvou tazích, porovnáváme buď dvě atomické hodnoty, nebo atomickou hodnotu a množinu. Případ, kdy bychom měli porovnávat dvě různé množiny, nemůže v našem případě nastat, protože desambiguace využívající lingvistických pravidel pracuje vždy s atomickými hodnotami a stochastické taggery používají totožné tagsety. Pro každou pozici v tagu tak dostaneme jeden ze tří výsledků: hodnoty se neliší, hodnota v jednom tagu je podmnožinou hodnoty v druhém tagu, anebo hodnoty se liší. Při porovnání dvou tagů pak mohou nastat tyto situace:

1. oba tagy se od sebe neliší;
2. existuje-li pozice, na které se hodnoty liší, řekneme, že se tagy liší;
3. nenastal případ 2 a alespoň jedna pozice v jednom tagu je podmnožinou pozice v druhém tagu -- potom řekneme, že celý první tag je podmnožinou druhého tagu (a naopak druhý tag je nadmnožinou prvního tagu).

Pro porovnání dvou tagů se bere v potaz pouze prvních dvanáct pozic. Pozice 13 a 14 jsou totiž rezervovány pro interní hodnoty využívané lingvistickými pravidly a na konečném výstupu obsahují vždy hodnotu -. Pozice 15 je vyhrazena pro vyznačení stylové varianty (hovorový tvar, knižní tvar, zkratka atd.) a často se na ní vyskytují chyby pocházející z morfologické analýzy, které desambiguační procedura není schopna napravit. Proto je i tato hodnota z porovnání vyloučena, aby nezkreslovala výsledky porovnání.
Porovnání tagů je ilustrováno v příkladu (1).

(1) a. VB-P---3P-AA--1 se neliší od VB-P---3P-AA--- (hodnoty 1 a - na patnácté pozici se neberou v úvahu);

b. NNIS4-----A---- se liší od NNIS1-----A---- (atomická hodnota 4 na páté pozici se nerovná atomické hodnotě 1);

c. VpYS----R-AA--- je nadmnožinou VpMS----R-AA--- (množina Y na třetí pozici v sobě obsahuje atomickou hodnotu M);

d. VB-P---3P-AA--1 se liší od VB-S---3P-AA--- (atomická hodnota P na čtvrté pozici se nerovná atomické hodnotě S, 1 a - na patnácté pozici se neberou v úvahu)

e. NNMXX-----A---8 se liší od NNIS1-----A---8 (atomická hodnota M na třetí pozici se nerovná atomické hodnotě N, množiny X na čtvrté a páté pozici už na porovnání nemají vliv)

f. PSXXXXP3------- je nadmnožinou PSMP1FP3------- (množiny X na třetí, čtvrté, páté a šesté pozici v sobě zahrnují atomické hodnoty M, P, 1 a F a na ostatních pozicích se tagy rovnají.

Teoreticky může nastat i případ, kdy jedna pozice v jednom tagu je obsažena v odpovídající pozici v druhém tagu a zároveň jiná pozice v prvním tagu je množinou obsahující odpovídající pozici z druhého tagu, ale prakticky tento případ nenastane, protože vždy porovnáváme buď dva tagy s atomickými hodnotami, nebo porovnáváme tag s atomickými hodnotami proti tagu obsahujícímu proměnné, anebo porovnáváme dva tagy s totožnými proměnnými.

Při vlastní evaluaci se za chybu považuje pouze případ, kdy se tagy liší. Pokud testovací korpus obsahuje obecné tagy a měřený korpus tagy atomické, pak nedokážeme změřit, kolika chyb se desambiguační procedura dopustila při snaze rozhodnout, která z hodnot obsažených v množině je správná. Pokud naopak testovací korpus obsahuje atomické tagy a měřený korpus tagy obecné, naměříme úspěšnost přesně.

5.2. Testovací korpus

Vzhledem k tomu, že se od r. 2007 změnila morfologická analýza, tagset i desambiguační metody, bylo by žádoucí vytvořit i nový ručně označkovaný korpus pro testovací účely. Bohužel však dosud žádný dostatečně velký testovací korpus nemáme k dispozici, a proto bylo rozhodnuto využít znovu testovacích dat z PDT, konkrétně tzv. d-test (testovací data určená pro vývoj) z dat pro t-rovinu obsahující cca 88.000 slov (vč. interpunkce). Hlavní nevýhoda tohoto testovacího korpusu je v tom, že používá obecné tagy, takže část chyb, kterých se dopustí modul lingvistických pravidel, zůstane neodhalena.

5.3. Normalizace dat

Aby byl výstup z nové desambiguace porovnatelný s daty z PDT a s výsledky stochastických taggerů, je třeba všechna data upravit do jednotné podoby odpovídající současnému způsobu tagování (viz oddíl 4). Šlo o tyto úpravy:

1. vypuštění čísla u zvratných zájmen *se*, *si*, *sebe*, *sobě*, *sebou*;
2. vypuštění osoby u příčestí;
3. vypuštění času u trpného příčestí;
4. vypuštění čísla u třetí osoby kondicionálu;
5. změna slovního druhu z adverbia na částici u vybraných slov.
6. odstranění vysvětlivek u lemmat (týká se výstupu z HMM a JH taggeru)

Naopak obecné i atomické tagy byly ponechány ve své podobě.

5.4. Výsledky nové evaluace

Při nové evaluaci jsme se snažili zopakovat všechna měření provedená v r. 2007. Oproti předchozí evaluaci jsme však měli k dispozici novější program morfologické analýzy, novější JH tagger i Morče, stejně tak jako novější lingvistická pravidla. Naopak naše verze HMM

taggeru byla starší, než byla použita při předchozí evaluaci. Výsledky samostatných stochastických taggerů (na tazích s proměnnými) jsou uvedeny v Tabulce 4.

tagger	úspěšnost
JH	91,72 %
HMM	92,61 %
Morče	94,94 %

Tabulka 4: Úspěšnost samostatných stochastických taggerů

V Tabulce 5 jsou uvedeny výsledky pro různé kombinace metod, tentokrát však byly použity tagy s atomickými hodnotami. Pro porovnání je v tabulce i výsledek taggeru Morče na atomických tazích.

metoda	úspěšnost
Morče	93,34 %
pravidla+Morče	94,57 %
SUBPOS+pravidla+Morče	94,80 %
sjednocení taggerů+pravidla+Morče	95,54 %

Tabulka 5: Úspěšnost kombinovaných metod na atomických tazích

Mohlo by se zdát, že od předchozí evaluace došlo k regresi (úspěšnost 95,68 % v r. 2007 oproti 95,54 %), ale je třeba si uvědomit, že při předchozí evaluaci se pracovalo s tagy obsahujícími proměnné, což mohlo výsledky vylepšit. Dalším handicapem pro desambiguaci atomických tagů je to, že závěrečnou fázi provádí stochastický tagger, který byl původně natrénován na odlišná data. Výsledky by tedy mohly být zlepšeny přetrénováním stochastického taggeru Morče.

6. Evaluace kvanta vykonané práce

Pro úplné posouzení úspěšnosti jednotlivých metod je také užitečné znát, jaké kvantum práce každá z nich vykoná. V Tabulce 6 je znázorněno, kolik tagů připadá na jedno slovo průměrně pro celý korpus, jaké je procento unikátních tagů (tzn. takových, které jsou již jednoznačně určeny) a kolik tagů připadá průměrně na jedno slovo, nepočítáme-li do průměru unikátní tagy. Jednotlivé sloupce ukazují hodnoty po provedení původní morfologické analýzy s tagy obsahujícími proměnné (morf(H)), po rozgenerování tagů na atomické hodnoty (morf(A)), po prvním kole pravidel (rules), pro desambiguaci provedené pomocí identifikace frazémů (frazrl) a po druhém kole pravidel s heuristikou (rulheu1).

	morf(H)	morf(A)	rules	frazrl	rulheu1

tag/form	3,88	12,81	2,71	2,70	2,17
formy s unikátními tagy	44,40 %	36,53 %	64,63 %	65,02 %	71,14 %
tag/form pro neunikátní tagy	6,18	19,60	5,84	5,85	5,04

Tabulka 6: Množství tagů v různých fázích zpracování

Z tabulky je vidět, že desambiguace tagů s atomickými hodnotami je mnohem pracnější než u obecných tagů, protože rozgenerováním tagů se jejich počet přibližně ztrojnásobí. Zároveň je vidět, že v modulu lingvistických pravidel udělá nejvíce práce první kolo bezpečných pravidel, která eliminují více než dvě třetiny tagů. Že to není na úkor přesnosti, je vidět v Tabulce 7, která zachycuje tzv. *recall* (tzn. procento forem, u kterých je v nabídce správný tag) v jednotlivých fázích desambiguace.

fáze	recall
morfologie	99,25 %
bezpečná pravidla	99,09 %
frazémový modul	99,07 %
pravidla s heuristikou	98,82 %

Tabulka 7: Recall po jednotlivých krocích desambiguační procedury

7. Závěry

V článku byl popsán proces tagování používaný v současné době pro značkování Českého národního korpusu. Tento proces byl evaluován vůči ručně anotovanému korpusu a výsledky byly porovnány. Zároveň byla kvantifikována pracnost desambiguace atomických tagů oproti tagům s proměnnými. Bylo také navrženo přetrénování stochastického taggeru, což by mohlo zlepšit celkovou úspěšnost desambiguace.

Literatura

Hajič J., 2004, *Disambiguation of Rich Inflection (Computational Morphology of Czech)*. Vol. 1. Karolinum Charles University Press, Praha.

Hajič J., Hajičová E., Panevová J., Sgall P., Pajas P., Štěpánek J., Havelka J., Mikulová M., 2006, Prague Dependency Treebank v2.0. CDROM, Linguistic Data Consortium, Philadelphia. Dokumentace též dostupná z WWW: <http://ufal.ms.mff.cuni.cz/pdt2.0>.

Hnátková M., 2002, Značkování frazémů a idiomů v Českém národním korpusu s pomocí Slovníku české frazeologie a idiomatiky. *Slovo a slovesnost*, 63, č. 2, 117-126.

Jelínek T., 2008, Nové značkování v Českém národním korpusu. In *Naše řeč*, 91, 1, 13-20.

Krbec P., 2005, *Language Modelling for Speech Recognition of Czech*. Disertační práce, MFF UK, Praha.

Květoň P., 2006, *Rule-based Morphological Disambiguation*. Disertační práce, MFF UK, Praha.

Petkevič V., 2006, Reliable Morphological Disambiguation of Czech: Rule-Based Approach is Necessary. In *Insight into the Slovak and Czech Corpus Linguistics*, ed M. Šimková, Veda, Bratislava, 26–44.

Spoustová D., Hajič J., Votrubec J., Krbec P. , Květoň P., 2007, The Best of Two Worlds: Cooperation of Statistical and Rule-Based Taggers for Czech. In *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing*. ACL 2007, Praha, 67-74.

Votrubec J., 2006, Morphological Tagging Based on Averaged Perceptron. In *WDS'06 Proceedings of Contributed Papers*, MFF UK, Praha, 191-195.