Anotace víceslovných výrazů v Pražském závislostním korpusu

Pavel Straňák, Eduard Bejček
{stranak,bejcek}@ufal.mff.cuni.cz

I. Motivace a úvod

V článku pojednáme o probíhajícím projektu *poloautomatické anotace víceslovných výrazů*. V prvém plánu jde o doplňování dodatečné informace k datům PDT 2.0, v dlouhodobějším pohledu by mělo jít o součást přiblížení tektogramatické vrstvy PDT k tomu, jak byla tektogramatická analýza ve Funkčním generativním popisu (FGP) navržena. Vysvětlíme, jakou roli v našem pojetí hraje ruční anotace a co je možno provést automaticky.

V Pražském závislostním korpusu verze 2.0 (dále jen PDT) jsou víceslovné výrazy anotovány jako skupina uzlů s (mnohdy arbitrárními) "syntaktickými" závislostmi mezi sebou.

Náš přístup předpokládá vytvoření jediného uzlu pro celý víceslovný výraz.

V externím slovníku můžeme navíc o jednotlivých lexémech udržovat dodatečné informace, které dosud v PDT nijak zachycovány (v těchto případech) nebyly.

V článku chceme ukázat, že anotace je dobře definovatelná úloha a je tedy proveditelná z hlediska konzistence anotátora se sebou samým i mezi anotátory. Ukážeme také, že automatická předanotace nám nejen nezhoršuje výsledky, ale pomáhá jak zlepšit konzistenci, tak zvýšit rychlost práce.

Definujeme pojmy *víceslovný lexém*, *lexie* a *pojmenovaná entita*, jak je budeme používat dále. Srovnáme své pojetí s dostupnými přístupy: v čem je náš přístup jiný, kde jdeme dál, co už naopak nerozlišujeme.

Současné pojetí víceslovných lexémů a pojmenovaných entit v PDT 2.0 a navrhované úpravy

Prague Markup Language (PML) je XML formát vyvinutý pro representaci korpusových dat a využitý v PDT 2.0. Data PDT jsou anotována na třech rovinách: morfologické, povrchově syntaktické a na rovině hloubkové syntaxe.
Každá rovina popisu je ve formátu PML zaznamenávána ve vlastním souboru, ale PML umožňuje také vytváření dalších souborů pro popis jevů, který chceme udržovat odděleně od jednotlivých rovin PDT. Této vlastnosti jsme využili a vytvořili jsme s-rovinu pro popis jevů souvisejících s lexikálním významem.

Stručně charakterizujeme s-rovinu PML, její vztah k ostatním rovinám PDT a k anotačnímu slovníku, výhody pro anotaci i užití s-roviny jako "dočasného úložiště" finálních anotací před jejich integrací do PDT.

Představíme funktory, technická lemmata a struktury užívané v PDT pro reprezentaci frazeologických a jiných komplexních jednotek a navrhneme způsob jejich modifikace.
Srovnáme stávající podobu zachycení víceslovných lexémů a entit v PDT s naší novou anotací. Upozorníme na některá hesla, která vyžadují pro správnou representaci obligatorní atributy. Např. "mít k dispozici" (proti "mít dispozici"; obligatorní a/aux.rf "k"), "co nevidět" (obligatorní negace), "vědět si rady" (nutný genitiv), "k pohledání" (jen jediný uzel, v PDT zcela bez relevantní anotace). Ukážeme, jak by vypadala t-rovina po integraci našich anotací.

II. SemLex

V této sekci stručně charakterizujeme anotační slovník SemLex. Jeho výchozí podoba je tvořena hesly extrahovanými z Eurovocu, Českého wordnetu a Slovníku české frazeologie a idiomatiky. Stručně popíšeme jednotlivé zdroje, informace, které obsahují a neobsahují, a co z nich používáme. Původní hesla jsme museli pro naše potřeby vyčistit, k "základnímu tvaru" přidat tvar lematizovaný, morfologické značky a následně i závislostní tektogramatickou strukturu hesla. Ukážeme, proč jsou námi přidané informace nezbytné.

III. Předanotace

Zde představíme metody použité pro všechny čtyři typy automatické předanotace. Vysvětlíme, čím je pro nás výhodná.
První typ vychází pouze z morfologie a provedla ho Milena Hnátková. Pro značkovánívíceslovných lexií použila Slovník české frazeologie a idiomatiky.
Druhým typem je předanotace před každým kolem manuálních anotací. Při níjsou v celé nové dávce vyhledány ty víceslovné lexie ze SemLexu, které již byly někdy použity pro anotaci.
Třetí typ předanotace se provádí ve chvíli, kdy anotátor otevře nový soubor. Pak jsou v něm vyhledány ty víceslovné lexie, které nově použil (a tedy přidal do slovníku) v tomto kole anotace.
Čtvrtý a nejjednodušší typ předanotace se aktivuje vždy, když anotátor označí část textu jako výskyt nové víceslovné lexie nebo pojmenované entity. V tu chvíli je prohledán zbytek otevřeného dokumentu a jsou shodně označkovány i ostatní výskyty stejné jednotky. V článku popíšeme, jak v tuto chvíli získáváme a do SemLexu ukládáme závislostní tektogramatické struktury víceslovné lexie. Toho je využito v 2. a 3. typu předanotace.
Ovšem i pro lexie označkované prvním typem předanotace se struktura ukládá, takže systém najde jak jejich další výskyty, tak místa, která jsou sice označena jako výskyty stejné lexie, ale mají odlišnou tektogramatickou strukturu. To znamená obvykle chybu předanotace.

IV. Manuální anotace

Na manuální anotaci pracují paralelně dva anotátoři. Popíšeme jejich činnost a hlavně nástroj, který jsme pro tento účel vyvinuli. Zdůrazníme jeho hlavní výhodu, totiž že uchovává informace o hloubkové, tektogramatické struktuře vět, přestože anotátorům zobrazuje jejich povrchové vyjádření (pro snazší orientaci v textu). Právě tektogramatická struktura je ve skutečnosti objektem anotace. (Označí-li například anotátor spojení "na každý pád", ve skutečnosti je anotována část tektogramatického stromu skládající se ze dvou uzlů: jeden s t-lemmatem "pád" a jemu podřazený uzel s t-lemmatem "každý" a odkazem na pomocný a-uzel "na".)

V. Vyhodnocení anotací

Ukážeme, kdy je kromě struktury nutno zapisovat i obligatorní atributy (viz příklady na konci části I.), že většinou to ovšem nutno není a stačí velmi jednoduchá závislostní struktura. Rozebereme úpravy, které zavádíme v metodologii pro případy, kdy je rozšířenstruktury o atribut nezbytné ("zaznamenat předložku/negaci/pád/číslo" na vybraném slovu v anotované instanci lexie). Rozebereme i případy, kdy stávající způsob anotace není optimální (např. koordinace jako "První a druhá světová válka") a navržené řešení (zobrazený text neodpovídá a je rozšířen o doplněné uzly).

Závěrem vyhodnotíme úspěšnost anotace pomocí shody mezi anotátory a také pomocí shody s automatickými předanotacemi. Ukážeme, že předanotace jsou výhodné a pomáhají manuální anotaci.