VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV POČÍTAČOVÉ GRAFIKY A MULTIMÉDIÍ

FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF COMPUTER GRAPHICS AND MULTIMEDIA

LZE TO ŘÍCI JINAK ANEB AUTOMATICKÉ HLEDÁNÍ PARAFRÁZÍ

DIPLOMOVÁ PRÁCE
MASTER'S THESIS

AUTOR PRÁCE Bc. LUBOMÍR OTRUSINA
AUTHOR

BRNO 2009

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV POČÍTAČOVÉ GRAFIKY A MULTIMÉDIÍ

FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF COMPUTER GRAPHICS AND MULTIMEDIA

LZE TO ŘÍCI JINAK ANEB AUTOMATICKÉ HLEDÁNÍ PARAFRÁZÍ

AUTOMATIC IDENTIFICATION OF PARAPHRASES

DIPLOMOVÁ PRÁCE
MASTER'S THESIS

AUTOR PRÁCE — Bc. LUBOMÍR OTRUSINA
AUTHOR

VEDOUCÍ PRÁCE — doc. RNDr. PAVEL SMRŽ, Ph.D.
SUPERVISOR

BRNO 2009

Abstrakt

Automatické získávání parafrází je důležitou úlohou v oblasti zpracování přirozeného jazyka. Uplatnění nalezne v systémech provádějících odpovídání na otázky, získávání informací nebo shrnutí dokumentů. Tato práce má za úkol seznámit čtenáře s problematikou získávání parafrází a následně vytvořit systém, který z volného textu parafráze získává. Práce nejprve vysvětlí hlavní pojmy v této oblasti, jako jsou parafráze nebo parafrázové vzory. Dále shrne přístupy k získávání parafrází z různých zdrojů. V další části je popsán návrh systému, který je zaměřen na získávání parafrází mezi dvěma pojmenovanými entitami. Na závěr jsou popsány metody vyhodnocování těchto systémů a je provedeno vyhodnocení našeho systému a jeho srovnání s podobnými systémy.

Abstract

Automatic paraphrase identification is an important task in natural language processing. Many systems use paraphrases for improve performance e.g. systems for question answering, information retrieval or document summarization. In this thesis, we explain basic concepts e.g. paraphrase or paraphrase pattern. Next we propose some methods for paraphrase discovery from various resources. Subsequently we propose an unsupervised method for discovering paraphrase from large plain text based on context and keywords between named entity pairs. In the end we explain evaluation metods in paraphrase discovery area and then we evaluate our system and compare it with similar systems.

Klíčová slova

parafráze, získávání parafrází, parafrázové vzory, pojmenované entity

Keywords

paraphrase, paraphrase identification, paraphrase pattern, named entities

Citace

Lubomír Otrusina: Lze to říci jinak aneb automatické hledání parafrází, diplomová práce, Brno, FIT VUT v Brně, 2009

Lze to říci jinak aneb automatické hledání parafrází

Prohlášení

Prohlašuji, že jsem tuto diplomovou práci vypracoval samostatně pod vedením pana doc. RNDr. Pavla Smrže, Ph.D.

. . . . . .
Lubomír Otrusina
24. května 2009

Poděkování

Děkuji doc. RNDr. Pavlu Smržovi, Ph.D. za hodnotné rady a odborné vedení během mé práce.

© Lubomír Otrusina, 2009.
Tato práce vznikla jako školní dílo na Vysokém učení technickém v Brně, Fakultě informačních technologií. Práce je chráněna autorským zákonem a její užití bez udělení oprávnění autorem je nezákonné, s výjimkou zákonem definovaných případů.

Obsah

1 Úvod **3**

2 Parafráze **5**
2.1 Typy parafrází . . . 5
2.1.1 Členění parafrází dle rozsahu . . . 6
2.1.2 Lexikální a syntaktické parafráze . . . 6
2.1.3 Orientované relace mezi parafrázemi . . . 6
2.2 Využití parafrází v oblasti zpracování přirozeného jazyka . . . 7
2.2.1 Systémy pro odpovídání na otázky . . . 7
2.2.2 Systémy pro získávání a extrakci znalostí . . . 7
2.2.3 Systémy pro shrnutí textu . . . 7
2.2.4 Systémy pro strojový překlad . . . 8

3 Pojmenované entity a jejich značkování **9**
3.1 Pojmenované entity . . . 9
3.2 Značkování pojmenovaných entit v češtině . . . 9
3.2.1 PDT 2.0 . . . 10
3.2.2 Další zdroje pro podporu značkování pojmenovaných entit . . . 11
3.3 Značkování pojmenovaných entit v angličtině . . . 11
3.3.1 SuperSense Tagger . . . 11

4 Metody získávání a klasifikace parafrází **14**
4.1 Získávání parafrází z paralelních či porovnatelných jednojazyčných korpusů 14
4.2 Získávání parafrází z prostých textových dat . . . 15
4.3 Získávání parafrází z paralelních dvojjazyčných korpusů . . . 15
4.4 Klasifikátory parafrází . . . 16

5 Návrh systému pro získávání parafrází **17**
5.1 Nástin algoritmu . . . 17
5.2 Použitá data . . . 17
5.2.1 Data pro češtinu . . . 18
5.2.2 Data pro angličtinu . . . 18
5.3 Extrahování dvojic pojmenovaných entit . . . 19
5.3.1 Formát COUPLES pro uložení dvojic pojmenovaných entit . . . 20
5.3.2 Experimenty . . . 21
5.4 Hledání charakteristických slov pro dvojice pojmenovaných entit . . . 22
5.4.1 Experimenty . . . 23
5.5 Vytváření shluků dvojic pojmenovaných entit . . . 23

5.5.1 Experimenty . . . 23
5.6 Hledání vztahů mezi shluky . . . 24
5.6.1 Formát PARAPHRASES pro uložení nalezených parafrází . . . 27
5.6.2 Experimenty . . . 27

6 Vyhodnocení a výsledky systému **29**
6.1 Metody vyhodnocování . . . 29
6.1.1 Systémy pro získávání parafrází . . . 29
6.1.2 Systémy pro klasifikaci parafrází . . . 32
6.1.3 Pracovní semináře týkající se parafrází . . . 33
6.2 Evaluace systému a diskuze výsledků . . . 34
6.2.1 Kritéria vyhodnocování . . . 34
6.2.2 Grafické rozhraní pro vyhodnocování . . . 35
6.2.3 Citlivostní analýza a výsledky systému . . . 37
6.2.4 Nalezení slabých míst systému . . . 41
6.2.5 Technické parametry systému . . . 42

7 Závěr **44**

A Přílohy **50**

Kapitola 1

Úvod

Zpracování přirozeného jazyka na počítačích s sebou přináší mnoho problémů. Mezi hlavní problémy patří fakt, že v přirozeném jazyce můžeme tu samou informaci vyjádřit několika různými způsoby. Jednou z možností, jak toho dosáhnout, jsou parafráze. Pro úspěšné vyřešení mnoha problémů v této oblasti, jakými jsou např. strojový překlad, získávání informací či odpovídání na otázky, je nutné umět variabilitu jazyka rozpoznat. Lepší zvládnutí práce s parafrázemi by pomohlo vylepšit funkčnost systémů, které získávájí informace z textu, provádí shrnutí dokumentů, odpovídají na otázky a mnoha dalších.

Pro zvládnutí práce s parafrázemi by stačilo, abychom měli vybudovánu bázi znalostí, která nám pokryje všechny možné aktuální případy parafrází v přirozeném jazyce. Manuální tvorba takové báze znalostí člověkem je téměř nemožná. Jednak z důvodu časové a finanční náročnosti, ale také proto, že člověk je tvor, který se může mýlit a nedokáže sepsat výčet všech parafrází. Jednoduše proto, že tento seznam je velmi rozsáhlý a také sám všechny parafráze nezná. Proto se v posledních letech zaměřilo mnoho výzkumníků na tvorbu automatických metod získávání parafrází. Vývoj těchto metod prodělal značný pokrok, nicméně doposud se ani nejlepším systémům nedaří tento problém plně zvládnout. Nově vznikající systémy jsou stále lepší a je zde stále častěji vidět snaha o použití největšího a nejdostupnějšího zdroje dat, internetu. Internet je perspektivním zdrojem dat nejen kvůli jeho dostupnosti, ale také proto, že pokrývá širokou doménu různých odvětví, což se o dnes běžně používaných paralelních či porovnatelných korpusech říci nedá.

Úkolem této práce je vytvořit systém, který umí získávat parafráze z velkých neanotovaných zdrojů dat, jakým je třeba internet. Systém by měl zvládat vyhledávat parafráze v anglickém i českém jazyce. Návrh systému byl inspirován podobnou prací Satoshi Sekineho, která se zabývá získáváním parafrází mezi kontexty dvou pojmenovaných entit. Tato práce navazuje na stejnojmenný semestrální projekt, z něhož byly převzaty některé kapitoly, týkající se zejména teorie k dané problematice.

V druhé kapitole provedeme vymezení některých základních pojmů z oblasti získávání parafrází. Rovněž v této kapitole popíšeme druhy parafrází a pokusíme se uvést některé příklady, které pomohou čtenáři dané pojmy lépe pochopit. V třetí kapitole pak vysvětlíme, co jsou to pojmenované entity a zmíníme zde problematiku jejich značkování. Rovněž zde popíšeme značkovače, které náš systém bude využívat. Ve čtvrté kapitole provedeme stručný popis zdrojů dat, které se dají pro získávání parafrází použít. Jedná se o paralelní a porovnatelné jednojazyčné korpusy, obyčejné textové korpusy a paralelní dvojjazyčné korpusy. V této kapitole se rovněž zmíníme o některých algoritmech či postupech používaných pro získávání parafrází či jejich klasifikaci. Pátá kapitola se zaměřuje na návrh vlastního systému. Jsou zde popsána data a principy, které byly při návrhu využity. V šesté kapitole

se zaměříme na způsoby vyhodnocování systémů, které získávají a klasifikují parafráze. Budou zde popsány používané metody vyhodnocování a také se zde zmíníme o nejpoužívanějších korpusech parafrází. Na závěr této kapitoly provedeme vyhodnocení našeho systému. V poslední kapitole je provedeno stručné shrnutí celé práce.

Kapitola 2

Parafráze

Na začátku kapitoly se pokusíme vysvětlit, co jsou to parafráze, parafrázové vzory a jaký je mezi nimi rozdíl. Dále bude uvedeno rozdělení parafrází z různých hledisek. Následně se pokusíme nastínit možnosti použití parafrází a jejich přínos v oblasti zpracování přirozeného jazyka. Jak uvidíme, schopnost rozpoznat parafráze je pro tuto oblast klíčová a umožňuje nám výrazně zlepšit výsledky některých algoritmů v této oblasti. Pokud budou v textu dvě věty nebo vzory parafrázemi, budeme to značit symbolem $\sim$.

2.1 Typy parafrází

Existuje několik definic *parafrází*, které se v literatuře často liší. V této práci se budeme držet definice převzané z [4], která říká, že parafráze jsou alternativním způsobem vyjádření téže informace. Parafráze se dají v širším měřítku chápat jako zobecnění synonym pro fráze či celé věty. To, jak kteří lidé vyjadřují tu samou informaci, závisí na jejich úrovni vědomostí v dané oblasti, stylu, upovídanosti a nebo osobních preferencích.

V mnoha aplikacích nám nestačí nalézt vhodné parafráze, ale potřebovali bychom spíše jejich obecnější formu. To lze provést např. tak, že v parafrázích nahradíme některá odpovídající si slova proměnnými. Jako příklad můžeme uvést parafaráze *Červený automobil je majetkem pana Nováka.* $\sim$ *Pan Novák vlastní červený automobil.* a jejich zobecněnou formu *X je majetkem Y* $\sim$ *Y vlastní X*, kde jsme některá slova nahradili proměnnými *X* a *Y*. Tuto zobecněnou formu parafrází budeme označovat jako *parafrázové vzory*. U parafrázových vzorů je dále možno zohledňovat situaci, kdy nějaká proměnná může nabývat hodnoty pouze určité kategorie např. místo, čas, osoba, atd. Příkladem takového parafrázového vzoru je např. *ORGANIZACE vznikla DATUM* $\sim$ *ORGANIZACE byla založena DATUM*, kde jsou proměnné pojmenovány podle příslušné kategorie, jejichž hodnot mohou nabývat. V kapitolách 5 a 6 budeme často pod pojmem parafráze myslet parafrázové vzory. U těchto vzorů budeme pak názvy proměnných psát anglicky.

V některých pracích jako jsou např. [7] nebo [18] autoři používají namísto pojmu parafráze pojem *inferenční pravidlo*. Oba pojmy se od sebe však mírně liší. Autoři zde chtějí použít parafráze pro aplikaci v systémech odpovídajících na otázky, a proto tento pojem zahrnuje i fráze, které nejsou parafrázemi, ale jsou potenciálně použitelné v těchto systémech. Příkladem takové relace může být např. *X způsobil újmu Y* $\sim$ *Y obviňuje X* nebo *X prodal automobil Y* $\sim$ *Y zalpatil za automobil X*.

Podle [8] mezi nejčastější způsoby tvorby parafrází patří použití synonym, změna slovního druhu některých slov ve větě, přeuspořádání věty, rozdělení věty na několik menších

vět a nebo použití různých definic, frází a příkladů.

2.1.1 Členění parafrází dle rozsahu

Parafráze můžeme na základě jejich délky rozdělit na frázové, větné a parafráze většího rozsahu. Frázové parafráze mívají rozsah jen několik málo slov, často to jsou parafráze obsahující pouze jediné slovo[1]. Příkladem těchto parafrází mohou být fráze *X je majetkem Y* $\sim$ *Y vlastní X*. Hledáním těchto parafrází se zabývá např. [14], [25] nebo [26]. Větné parafráze svým rozsahem pokrývají celé věty či souvětí, jejich získáváním a tvorbou se zabývá např. [5] nebo [12]. Příkladem větných parafrází mohou být věty *Jsou si podobní jako vejce vejci.* $\sim$ *Jsou jako dvojčata.* Parafráze většího rozsahu se týkají odstavců, paragrafů nebo i celých dokumentů. Uplatnění systémů zabývajících se hledáním či tvorbou těchto parafrází je hlavně při odhalovaní plagiátorství a generování podobných dokumentů. Těmito parafrázemi se zabývá např. [23] nebo [28].

Systémy pro získávání parafrází se liší v závislosti na tom, jaké parafráze se snažíme nalézt. Systémy, které hledají parafráze nad celými dokumenty, používají odlišné algoritmy než systémy, které hledají parafráze mezi větami. S tímto omezením je potřeba počítat a zaměřit se na požadovaný typ parafrází. Pokud nebude uvedeno jinak, budeme v této práci jako parafráze chápat frázové parafráze.

2.1.2 Lexikální a syntaktické parafráze

Podle struktury či obsahu věty můžeme parafráze dále dělit na lexikální a syntaktické. Lexikální parafráze vznikají záměnou určité části textu za jinou. Přičemž není důležité, jestli původní text obsahoval větší, menší či stejný počet slov. Příkladem lexikálních parafrází mohou být např. věty *Spěchal jsem, proto jsem se rozběhl.* $\sim$ *Spěchal jsem, proto jsem začal utíkat.* Syntaktické parafráze vznikají změnou struktury věty při zachování důležitých slov ve větě. Jedná se často o různá přeuspořádání slov, či změnu trpného rodu na činný atd. Příkladem takových parafrází mohou být věty *Spěchal jsem, proto jsem začal utíkat.* $\sim$ *Začal jsem utíkat, protože jsem spěchal.* Nejčastěji se však setkáme s kombinací obou typů parafrází, kdy mezi větami je rozdíl jak v jejich struktuře, tak v jejich lexikálním složení.

2.1.3 Orientované relace mezi parafrázemi

Parafráze je možno chápat jako ekvivalentní textové řetězce. Tedy tam, kde můžeme použít jeden, lze použít i druhý. Často se však v běžném životě setkáme s případy, kdy bychom potřebovali spíše orientovanou relaci mezi dvěmi textovými řetězci, které mohou vyjadřovat stejnou informaci. Takové rozšíření je navrženo v [10] a autor jej nazývá *entailment relations*. My se budeme držet označení orientované relace mezi parafrázemi. Jedná se o relaci mezi textem T a jazykovým výrazem, který je uvažován jako hypotéza H. H je důsledek T, pokud význam H interpretován v kontextu T, může být odvozen z významu T. Tuto skutečnost značíme $T \rightarrow H$. Příkladem takové relace může být např. *X jí rád Y* $\rightarrow$ *X má rád Y*, kde $T =$ *X jí rád Y* a $H =$ *X má rád Y*. Vidíme, že pokud dosadíme H za T, dá se význam H odvodit z významu T. Pokud ovšem provedeme dosazení T za H, pravidlo neplatí, protože některé věci, které má X rád, nejsou k snědku. Tyto relace jsou proto ideální k vyjádření orientovaného vztahu mezi dvěmi frázemi. Získáváním těchto relací se zabývá např. [27].

V [7] je popsán algoritmus, který se snaží z dvojic parafrázových vzorů vytvořit orientované relace. Využívá k tomu rozšířenou Harrisovu distribuční hypotézu, která říká, že

[1] v takovém případě se jedná o synonyma

pokud se dvě parafráze vyskytují ve stejném kontextu a první se vyskytuje výrazně častěji, pak druhá implikuje první.

2.2 Využití parafrází v oblasti zpracování přirozeného jazyka

Parafráze mají v oblasti zpracování přirozeného jazyka velmi zásadní význam. Jejich využití nalezneme v systémech pro odpovídání na otázky, systémech pro získávání informací, systémech pro shrnutí textu a mnoha dalších odvětvích zpracování přirozeného jazyka i umělé inteligence obecně.

2.2.1 Systémy pro odpovídání na otázky

Parafráze pomáhájí v oblasti odpovídání na otázky řešit problém, kde je hledaná odpověď v prohledávaném dokumentu formulována jinak, než příslušná otázka. Pokud by systém uměl poznat v textu parafráze k odpovědi, mohl by identifikovat i odpovědi, které neobsahují stejné formulace jako otázka. Příkladem může být systém, který dostane za úkol odpovědět na otázku *Kdo napsal Hamleta?*. Uvažme, že systém hledá odpověď v datech, ve kterých ovšem nikde nejsou dána do souvislosti slova *napsat* a *Hamlet*. Prohledávaná data ovšem obsahují informaci *William Shakespeare je autorem díla Hamlet.*. Pokud by odpověď hledal člověk, nebude mít patrně problém určit, že Hamleta napsal William Shakespeare. Systém, který ovšem neví, že fráze *napsal* a *je autorem* jsou parafrázemi, nemůže správně odpovědět na zadanou otázku.

2.2.2 Systémy pro získávání a extrakci znalostí

Systémy pro získávání informací jsou systémy, které se snaží pro zadaný dotaz nalézt v textu co nejlepší odpověď. Jsou podobné systémům pro odpovídání na otázky. Pokud bychom zavedli do těchto systémů znalost parafrází, mohl by systém vstupní dotaz nahradit dotazem, který kromě původního dotazu obsahuje i jeho parafráze, a tím pádem by se zvětšila šance, že nalezneme správnou odpověď. Některé současné systémy, které neumí pracovat s parafrázemi, tento problém řeší tak, že k nalezeným relevantním odpovědím najdou odpovědi podobné a ty potom zařadí jako alternativní odpovědi, což nemusí být vždy správné.

Parafráze nachází rovněž uplatnění v systémech pro extrakci informací z textu. Tyto systémy většinou extrahují informace z jisté domény z nestrukturovaných strojově čitelných dat. Zavedením parafrází do těchto systémů, umožní systému nalézt potřebné informace i pokud se nevyskytují ve formě, kterou systém umí rozpoznat. Pokud systém zjistí, že některé informace jsou parafázemi k informacím, které umí zpracovat, je pak získání takovýchto informací snadnější.

2.2.3 Systémy pro shrnutí textu

Dalším odvětvím, kde se uplatní parafráze, je shrnutí textu z jednoho či více dokumentů. Systémy, které provádějí shrnutí textu, mají uplatnění např. při tvorbě abstraktů. Využití parafrází v těchto systémech je dvojí. Jednak slouží ke generování nových vět, které jsou parafrázemi vět obsažených v některém ze vstupních dokumentů. Při takovém generování umožňuje zavedení parafrází generovat mnohem rozmanitější věty, které mají i vyšší úroveň jazyka. Shrnutí napsané člověkem je totiž mnohem víc než jen pár vybraných vět z článku, často jsou to parafráze nebo spojení frází z více vět do jedné. V druhém případě jsou

parafráze používány k tomu, aby systém byl schopen rozpoznat u více dokumentů věty s podobným významem a nezahrnul tak do shrnutí některou informaci vícekrát.

2.2.4 Systémy pro strojový překlad

Mezi další aplikace parafrází v oblasti zpracování přirozeného jazyka patří např. využití ve strojovém překladu. Zde se parafráze používají především pro vyhodnocování a porovnávání systémů pro strojový překlad. Různé systémy pro strojový překlad totiž mohou překlad provést různě a přitom oba překlady mohou být správné. Použitím parafrází jsme pak schopni poznat, které překlady jsou správné a které ne. Ve strojovém překladu jsme rovněž pomocí parafrází schopni získat mnohem vyšší kvalitu textu.

Kapitola 3

Pojmenované entity a jejich značkování

V této kapitole se pokusíme vysvětlit, co jsou pojmenované entity a jak lze pomocí specializovaných nástrojů provádět jejich značkování. Zjistíme, že v obou jazycích je situace odlišná, a to zejména kvůli nedostupnosti kvalitních nástrojů pro češtinu. Vysvětlíme zde také postup jak lze tento nedostatek v češtině alespoň částečně kompenzovat.

3.1 Pojmenované entity

Termín *pojmenovaná entita* je překladem anglického termínu *named entity*. Podle [31] můžeme za pojmenované entity označit výrazy, které nemají apelativní význam. Jedná se o slova či slovní spojení, která v textu vystupují jako jména osob (Barack Obama), geografcké názvy (La Rochelle), jména produktů (AutoCAD 2009), názvy organizací (ŠKODA HOLDING a. s.) a institucí (Komora logistických auditorů), ale také jako časové údaje (1995) apod. S pojmem pojmenované entity úzce souvisí pojem *vlastní jména*. Vlastní jména jsou většinou charakteru podstatných jmen, která označují konkrétní skutečnost (osoba, zvíře, věc, jednotlivina atd.). Vlastní jména jsou charakteristická také tím, že začínají velkým počátečním písmenem (Paříž, Ghándí, Na Nivách atd.). Pojmenované entity jsou pak chápány jako nadmnožina vlatních jmen, obohacená o další výrazy jednoznačně odkazující k objektům (rodiště J. A. Komenského), osobám (americký prezident), datům (studená válka) atd. Pojmenované entity se od vlastních jmen liší také tím, že o pojmenovaných entitách mluvíme v oblasti zpracování přirozeného jazyka, kdežto o vlastních jménech v lingvistické oblasti. Pro značkování pojmenovaných entit se používají speciální nástroje, nazvané značkovače pojmenovaných entit[1] (dále jen značkovače).

3.2 Značkování pojmenovaných entit v češtině

Pro značkování pojmenovaných entit v češtině zatím neexistují žádné uspokojivě fungující nástroje. Vytvářením takového nástroje se zabývá projekt Informační společnosti Grantové agentury Akademie věd České republiky nazvaný Integrace jazykových zdrojů za účelem extrakce informací z přirozených textů. Jelikož plánované dokončení projektu má být nejdříve v roce 2009, nelze tento nástroj zatím použít. Jako alternativu ke značkovači pojmenova-

[1] Named Entity Tagger

ných entit lze použít softwarové nástroje *Pražského závislostního korpusu (PDT)*, případně další podpůrné informace jako jsou např. seznamy jmen, názvů měst, ulic atd.

3.2.1 PDT 2.0

PDT [15] je projekt pro ruční anotaci velkého množství českých textů na několika úrovních (morfologická, sémantická, syntaktická, pragmatická). PDT 2.0 obsahuje 2 milióny slov s provázanými anotacemi na úrovni morfologie (2 milióny slov), povrchové syntaxe (1,5 mil. slov) a hloubkové syntaxe a sémantiky (0,8 mil. slov). Data pochází z různých novin a časopisů (Lidové noviny, Mladá fronta Dnes, Českomoravský Profit, Vesmír). Obsahuje také softwarové nástroje pro prohledávání korpusu, automatickou anotaci dat na různých úrovních a jazykovou analýzu. Součástí morfologické anotace je také označení vlastních jmen a jejich zařazení do kategorií, což lze do značné míry chápat jako označkování pojmenovaných entit. Toto označkování je bohužel prováděno již na morfologické úrovni a zahrnuje pouze jednoslovná vlastní jména. Tento fakt může být značně limitující. Kvůli neexistenci jiného značkovače pojmenovaných entit pro češtinu se s ním však spokojíme.

Morfologická anotace textu pomocí nástrojů z PDT probíhá tak, že je každému slovu přiřazena morfologická značka a morfologické lemma. Morfologická značka obsahuje 15 znaků (poziční systém navržený pro PDT), které vyjadřují slovní druh a různé morfologické kategorie (pád, číslo, čas atd.). Morfologické lemma se skládá ze dvou částí. První část obsahuje vlastní lemma. Druhá část, která není povinná, obsahuje dodatečné přípony. Přípony se dělí do 4 kategorií, které se mohou v morfologickém lemmatu vyskytovat, ale také nemusejí. Navíc se některé kategorie přípon mohou vyskytovat i vícekrát. Zájemce o podrobnější informace o struktuře morfologického lemmatu a značky lze odkázat na [16]. Označíme-li kategorie *K1* až *K4*, pak morfologické lemma můžeme zapsat ve tvaru:

```
lemma_:K1;_K2,_K3_^(K4).
```

Jednotlivé významy kategorií jsou vysvětleny v tabulce 3.1.

Kategorie	**Význam kategorie**
K1	informace o vidu; informace, zda se jedná o zkratku
K2	typ vlastního jména
K3	stylový příznak (hovorová čeština, slangový výraz, cizí slovo atd.)
K4	komentáře různého typu (význam slova, derivační informace atd.)

Tabulka 3.1: Tabulka popisující kategorie přípon morfologického lemmatu v PDT 2.0.

Nejzajímavější je pro nás kategorie *K2*, která označuje typ vlastního jména, jenž může nabývat některé z hodnot uvedených v tabulce 3.2. Tyto kategorie pak můžeme brát jako základní kategorie pojmenovaných entit.

Kvůli tomu, že přiřazování těchto kategorií probíhá již na morfologické úrovni, není výsledné přiřazení značek moc spolehlivé. Dochází k častým chybám způsobených tím, že každé slovo je analyzováno samostatně. U víceslovných vlastních jmen dochází k případům, kdy část vlastního jména je správně označkována, ale zbytek již ne. Tuto situaci ilustruje příklad označkování vlastního jména *Fakulta architektury ČVUT*. Pro přehlednost nejsou uváděny morfologické značky, ale pouze morfologická lemmata.

```
fakulta architektura ČVUT-1_:B_;K_^(České_vysoké_učení_technické)
```

Mezi další problémy patří fakt, že i když značkovač vyhodnotí každé slovo víceslovného vlastního jména jako vlastní jméno, může každému slovu přiřadit jinou kategorii, některým slovům dokonce více kategorií. To komplikuje následnou práci s takto označkovanými vlastními jmény, protože zpravidla chceme celému vlastnímu jménu přiřadit pouze jedinou značku.

Podle [31] další nevýhodou tohoto ohodnocování vlastních jmen je nekompletní skupina kategorií vlastních jmen. Systém počítá s kategorií barva, ale ne již s takovými kategoriemi jako tvar. Navíc některé kategorie nejsou při ruční anotaci striktně dodržovány, např. slova červený, fialový jsou označkovány jako vlastní jména kategorie barva, ale slovo modrý již ne.

Všechny tyto aspekty je potřeba brát v úvahu při používání nástrojů PDT 2.0 coby značkovače pojmenovaných entit.

3.2.2 Další zdroje pro podporu značkování pojmenovaných entit

Mezi další zdroje, které by bylo možno použít, patří výčtové seznamy pojmenovaných entit. Jedná se např. o seznamy názvů ulic, náměstí, měst, křestních jmen, příjmení atd. Takovéto seznamy lze nalézt např. na školním serveru *merlin* v adresáři */mnt/minerva1/nlp/projects/ner* nebo */mnt/minerva1/nlp/projects/ner2*. Další podobné seznamy lze získat např. na stránkách *Českého statistického úřadu*[2]. Kvůli jejich neúplnosti a problematickému zpracování je v našem systému používat nebudeme.

3.3 Značkování pojmenovaných entit v angličtině

V angličtině je problematika značkování pojmenovaných entit jednodušší kvůli existenci mnoha dostupných značkovačů. Z volně dostupných zmiňme např. nástroj *GATE*[3] nebo *SuperSense Tagger*[4], který byl použit při označkování dat pro náš systém.

3.3.1 SuperSense Tagger

Jedná se o sémantický značkovač, jehož bližší popis lze nalézt v [20]. Značkovač používá skryté Marokovovy modely, algoritmus perceptron a je založen na kategoriích z WordNetu[5]. Počet kategorií, které lze s pomocí tohoto značkovače rozpoznat, je velký. Je nutno podotknout, že SuperSense Tagger nezahrnuje mezi pojmenované entity pouze vlastní jména, ale také některá obecná jména, což ovšem funkčnosti našeho systému nebude vadit. Kategorie, které umí značkovač rozpoznat, byly převzaty z [29] a lze je nalézt v tabulce 3.3. Názvy kategorií jsou ponechány v angličtině.

[2] http://www.czso.cz/
[3] http://gate.ac.uk/
[4] http://sourceforge.net/projects/supersensetag/
[5] http://wordnet.princeton.edu/

Hodnota	**Popis**	**Příklad**
Y	křestní jméno (dříve jako defualtní hodnota)	Ivan, Jakub, Marie atd.
S	příjmení, rodné jméno	Kasparov, Tošovský, Voskovec atd.
E	příslušník národa, obyvatel území	Pražák, Čech, Američan atd.
G	geografické jméno	Vídeň, Ankara, Blansko atd.
K	společnost, organizace, instituce	Ferona, Sigma, Opel atd.
R	produkt	DOS, Lanza, lego atd.
m	ostatní vlastní jména: jména stadionů, dolů, partyzánských základen atd.	Prix, Cup, Leviathan atd.
H	chemie	molybden, panthenol, dibenzofuran atd.
U	lékařství	dilatátor, kancerofobie, osteoartróza atd.
L	přírodní vědy	mtDNA, hlístice, octomilka atd.
j	právo	aequo, obedience, půhon atd.
g	technologie obecně	stereovize, turboalternátor, fázoměr atd.
c	výpočetní technika a elektronika	HTML, CD, konduktivita atd.
y	záliby, volný čas, cestování	MTV, CD, CHKO atd.
b	ekonomie, finance	USD, euro, ČNB atd.
u	kultura, vzdělávání, umění, ostatní vědy	agitato, Beatles, friska atd.
w	sport	NHL, FC, Cup atd.
p	politika, vláda, armáda	ODS, ČSSD, EU atd.
z	ekologie, životní prostředí	komenzalizmus, CHKO, transekt
o	barvy	hnědobílý, červený, fialový, atd.

Tabulka 3.2: Tabulka s kategoriemi vlastních jmen podle PDT 2.0. Příklady byly vybrány ze zpracovaných dat.

Typ entit	**Popis**
pojmenované entity	Person, Facility, Organization, GPE, Location, Nationality, Product, Event, Work of Art, Law, Language, Contact-Info
nominální entity	Person, Facility, Organization, GPE, Product, Plant, Animal, Substance, Disease, Game
numerické entity	Date, Time, Percent, Money, Quantity, Ordinal, Cardinal

Tabulka 3.3: Tabulka s kategoriemi entit, které lze s pomocí SuperSense Taggeru rozpoznat.

Kapitola 4

Metody získávání a klasifikace parafrází

V této kapitole se pokusíme uvést přehled metod používaných pro získávání a klasifikaci parafrází. Tyto metody se dají rozdělit na tři skupiny z hlediska zdrojů použitých dat. Prvním zdrojem jsou paralelní a porovnatelné jednojazyčné korpusy, což jsou korpusy, které obsahují většinou stejné informace popsané různou formou. Druhou skupinou jsou metody získávající parafráze z obyčejných textových dat, která jsou běžně dostupná a dnes nejrozšířenější. Poslední skupinu tvoří paralelní dvojjazyčné korpusy.

4.1 Získávání parafrází z paralelních či porovnatelných jednojazyčných korpusů

Paralelní jednojazyčné korpusy jsou korpusy, které obsahují zarovnaný text, tzn. odpovídající si věty jsou zarovnány. Tyto korpusy jsou vysoce redundantní, protože většinu informací obsahují dvakrát. Takové korpusy se v reálném světě vyskytují zřídka. Jejich tvorba totiž, narozdíl od paralelních vícejazyčných korpusů, není příliš běžná. Častěji se setkáme s porovnatelnými korpusy, které neobsahují přímo zarovnané věty, ale některé jejich věty lze zarovnat. Příkladem porovnatelných korpusů může být korpus vícenásobných překladů z cizího jazyka nebo korpus novinových článků popisujících stejné události.

U porovnatelných korpusů vícenásobných překladů z cizího jazyka se často jedná o prózu, kde bývá použit zastaralý jazyk, bývá často chráněna copyrightem a mívá úzké doménové zaměření. Při získávání parafrází z korpusu vícenásobných překladů cizojazyčných děl se často používají metody podobné těm používaným při strojovém překladu. Nelze použít metody používané přímo pro strojový překlad, protože zarovnání v těchto korpusech není zdaleka tak dobré[1] jako u dvojjazyčných korpusů. Využívá se zde skutečnosti, že některé věty mohou být přeloženy podobně, nebo alespoň obsahují některé společné informace jako jsou jména, místa, časy, atd. Tento fakt na jednu stranu ulehčuje identifikaci podobné věty, ale na druhou stranu znemožňuje použití technik používaných při strojovém překladu. Příklad systému, který používá cizojazyčné překlady lze nalézt např. v [6] nebo [21].

U korpusu vytvořeného z novinových článků je potřeba vyřešit identifikaci odpovídajících si článků. K tomu slouží různé podpůrné informace jakými jsou např. informace o datu článku. Často se zde také využívá faktu, že novinové články se snaží v prvních několika

[1] kvalita zarovnání se měří pomocí Alignment Error Rate, která se běžně používá ve strojovém překladu

větách shrnout obsah celého článku, a proto se první věty článku mezi sebou porovnávají na překryv. U tohoto zdroje dat se rovněž využívá informací, jako jsou např. data, čísla, místa, jména, atd. Tyto informace jsou získány pomocí značkovače. Když jsou nalezeny odpovídající si věty, je provedena extrakce příslušných částí vět. U novinových článků se často také využívá předpokladu, že parafráze v nich obsažené tvoří svým rozsahem právě jednu větu. Systémy pracující s novinovými články je možné nalézt např. v [5], [9] nebo [26].

V práci [8] byl použit poněkud odlišný zdroj dat, který ovšem rovněž patří do kategorie porovnatelných korpusů. Autorka použila sadu odpovědí studentů na zadávané otázky. Vycházela z předpokladu, že mezi otázkami a odpověďmi bude možno nalézt parafráze.

4.2 Získávání parafrází z prostých textových dat

Dalším zdrojem dat mohou být prostá textová data (např. webové stránky), která mají tu výhodu, že jsou snáze dostupná a mohou mít širší doménové pokrytí. Mezi jejich nevýhody patří skutečnost, že získání parafrází z těchto dat je obtížnější, protože reálná data obsahují často šum a nedá se předem určit, které dvě věty mohou být kandidáty na parafráze, což u porovnatelných korpusů není takový problém. Metody pracující s těmito daty používají algoritmy shlukování, které často nepotřebují žádné anotace či trénovací data. Místo globálního kontextu věty se zde pracuje spíše s lokálním kontextem. Mezi důležité vlastnosti potřebné k rozpoznávání parafrází patří např. sekvence slov, sekvence POS[2] značek nebo části syntaktického stromu věty.

Metoda získávání parafrází použitá v [14] nebo [25] pracuje na principu identifikace pojmenovaných entit, které se dají získat pomocí značkovače. Ze získaných značek se následně utvoří dvojice. Parafráze se pak hledají mezi kontexty těchto dvojic. Příkladem takové dvojice může být např. *OSOBA je zaměstnancem ORGANIZACE*. Tyto dvojice se na základě podobnosti kontextů shlukují. Na závěr se mezi jednotlivými shluky hledají podobnosti na základě stejných instancí entit ve shlucích. Tohoto principu využívá i náš systém pro získávání parafrází.

Jiný princip je využit v [18], kde se pracuje s cestami v syntaktických stromech. Algoritmus se snaží nalézt podobné cesty ve stromech. Čím více stejných vlastností cesty sdílí, tím více jsou si podobné. Nejpodobnější cesty jsou pak prohlášeny za parafráze. Využívá se zde rozšířené Harrisovy distribuční hypotézy, která předpokládá, že cesty, které se vyskytují v podobných kontextech, mají podobný význam.

4.3 Získávání parafrází z paralelních dvojjazyčných korpusů

Paralelní dvojjazyčné korpusy jsou poměrně snadno dostupné zdroje. Pro získávání parafrází se ovšem často nepoužívají. Doposud nebylo publikováno mnoho metod, které dokáží získávat parafráze z těchto korpusů. Tyto korpusy jsou použity např. v [4]. Autoři zde používají algoritmus, který je z korpusu schopen získávat parafráze v obou jazycích současně. Princip algoritmu je založen na nalezení všech možných překladů určité fráze. Ke každé frázi je pak možno nalézt až několik různých parafrází v druhém jazyce. Důležitým faktem, který se v [4] podařilo prokázat, je, že kvalita nalezených parafrází je závislá na kvalitě zarovnání takových zdrojů. V tomto experimentu autoři srovnávali kvalitu nalezených

[2] Part-of-Speech

parafrází u automaticky a ručně zarovnaných vět. U ručního zarovnání vět pak dosahovaly nalezené parafráze větší kvality. Lze předpokládat, že při vylepšení kvality zarovnání získáme i kvalitnější parafráze.

4.4 Klasifikátory parafrází

Za klasifikátory parafrází považujeme algoritmy, které jsou schopny určit, zda dvě věty jsou parafrázemi či nikoliv. Mezi nejčastěji používané klasifikátory patří např. Support Vector Machines [9] či rozhodovací stromy [30]. Některé klasifikátory požívají k natrénování mnoho různých vlastností, jako např. míru podobnosti vět, morfologické varianty, n-gramy, délky vět, počty sdílených slov mezi větami atd. Většinou je potřeba trénovací data ručně anotovat, což je časově náročné.

Zajímavou prací je [30], kde jsou dvě věty před vlastní klasifikací nejprve převedeny do tzv. základní formy. Algoritmus spoléhá na to, že dvě parafráze převedené do základní formy si budou mnohem více podobné než dvě věty, které nejsou parafrázemi. Mezi operace, které jsou aplikovány na větu pro její převedení do základní formy, patří např. nahrazení pojmenovaných entit jejich značkami, změna pasivního tvaru na aktivní či explicitní označení budoucího času. Po převedení obou vět do základní formy se určí podobnost vět, a pokud tato hodnota bude menší než určitý práh, jsou věty prohlášeny za parafráze.

Jiné použití klasifikátoru je uvedeno v [24], kde autoři každé dvě věty, jenž představují potenciální parafráze, rozdělí na několik malých informačních částí, které se pak k sobě snaží navzájem jednoznačně přiřadit. Pokud se podaří části navzájem přiřadit, jsou obě věty považovány za parafráze. Pokud některá část přebývá, použije se na ni klasifikátor významnosti, který rozhodne, jestli je část významná či nikoliv. Pokud část není významná, věty jsou prohlášeny za parafráze, v opačném případě věty nejsou parafráze.

Kapitola 5

Návrh systému pro získávání parafrází

Systém, který je součástí této práce, vychází převážně z [25] a některé myšlenky jsou převzaty z [17]. Systém je zaměřen na hledání parafrází mezi kontexty dvou pojmenovaných entit. Nejprve se pokusíme stručně popsat celý algoritmus a použitá data, následně provedeme jeho detailnější rozbor po jednotlivých krocích.

5.1 Nástin algoritmu

Systém zpracuje vstupní data pomocí vhodného značkovače, který označkuje pojmenované entity v datech jejich kategoriemi. Následně extrahuje uspořádané dvojice značek pojmenovaných entit (dále jen dvojice značek) spolu s jejich kontextem. Kontextem jsou chápána slova, která se nachází mezi oběma značkami. Kontext spolu s příslušnou dvojicí značek nazveme frází, dvojici konkrétních hodnot pojmenovaných entit pak instancí. Množina všech frází, které jsou tvořeny stejnou dvojicí značek, pak tvoří doménu. Inverzní doména k doméně je taková, která je charakterizována dvojicí značek v opačném pořadí. V některých případech nebudeme pracovat s instancemi, ale s lemmaty instance. Lemma instance vznikne tak, že jsou všechna slova instance nahrazena jejich lemmaty. Pro lepší pochopení pojmů fráze, doména, instance atd. jsou v tabulce 5.1 uvedeny příklady.

V češtině mohou být lemmata, která jsou přiřazena pomocí nástrojů z PDT 2.0, někdy velmi dlouhá. Jejich vypisování by bylo nepřehledné, a proto od takovýchto lemmat odstraníme všechny přípony. V některých případech (obzvláště v obrázcích) pak budeme lemmata psát pro lepší přehlednost velkými písmeny.

Pro každou frázi ve všech doménách nalezneme charakteristické slovo, které určitým způsobem nejvýstižněji popisuje daný kontext. Toto slovo určíme pomocí metriky *tf–idf*[1]. Všechny fráze se stejným charakteristickým slovem v dané doméně umístíme do jednoho shluku. Na závěr se snažíme pomocí instancí dvojic nalézt mezi shluky vztahy, které vyjadřují parafráze. Celý algoritmus je znázorněn na obrázku 5.1.

5.2 Použitá data

Neméně důležitým aspektem této práce jsou použitá data. Ta by měla mít širší doménové pokrytí a měla by obsahovat dostatek pojmenovaných entit, aby výše popsaný algoritmus

[1] term frequency/inverse document frequency

Pojem	Příklad
text	Moskvu navštívil americký ministr Warren Christopher
katagorie pojmenovaných entit	{LOCATION, PERSON}
dvojice značek	<LOCATION, PERSON>
fráze	LOCATION navštívil americký ministr PERSON
dvojice pojmenovaných entit	<Moskvu, Warren Christopher>
kontext	navštívil ministr zahraničí
označení domény	LOCATION-PERSON
označení inverzní domény	PERSON-LOCATION
instance dvojice	<Moskvu, Warren Christopher>
lemmata instance dvojice	<Moskva_;G, Warren_;G_;S Christopher-1_;S>

Tabulka 5.1: Příklady vysvětlující pojmy používané při návrhu systému.

nalezl dostatečné množství kvalitních parafrází.

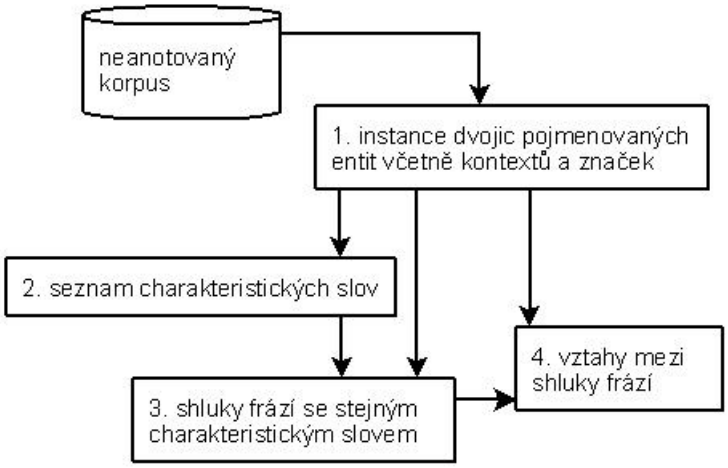
5.2.1 Data pro češtinu

Pro češtinu bylo plánováno použít korpusu novinových článků. K jejich stahování byl použit nástroj Heritrix[2], který provádí stahování celých webů. Jako zdroje dat byly použity servery *novinky.cz* a *zpravy.idnes.cz*. Stahování a následné čištění dat z těchto serverů ovšem zabralo značné množství času, proto se povedlo zpracovat jen velmi malou část dat, která pro účely automatického hledání parafrází není příliš vhodná. Byl proto zvolen jiný zdroj dat, kterým se stal velký korpus češtiny poskytnutý FI MU[3], který obsahuje kolem 600 milionů slovních tvarů. Tento korpus byl zpracován pomocí anotačních nástrojů z PDT 2.0, které jsou popsány v [15].

5.2.2 Data pro angličtinu

Pro angličtinu byla jako zdrojová data vybrána data projektu Semantically Annotated Snapshot of the English Wikipedia (SW v.1), který je popsán v [29]. Tento korpus byl vytvořen z anglické verze wikipedie[4] pomocí několika volně dostupných nástrojů z oblasti zpracování přirozeného jazyka. Mezi použité nástroje patří např. The dependency parser[5] nebo SuperSense Tagger popsaný v kapitole 3.3.1. Jedná se o verzi wikipedie z 4. 11. 2006, která obsahuje 1490688 záznamů. Celý korpus je rozdělen do 3000 souborů, kde každý obsahuje zhruba 500 záznamů. Autoři korpusu upozorňují, že všechny značky (slovní druhy, sémantické a závislostní vztahy) byly získány automatickým procesem a mohou proto obsahovat chyby a že jim nejsou známy přesné úspěšnosti přiřazení jednotlivých značek. Kvůli značnému množství dat, a tudíž i velké paměťové náročnosti, je pro hledání parafrází použito pouze prvních 200 souborů tohoto korpusu, dle abecedního pořadí.

[2] http://crawler.archive.org/
[3] Fakulta informatiky Masarykovy univerzity
[4] http://en.wikipedia.org
[5] http://desr.sourceforge.net/doc



Obrázek 5.1: Diagram popisující algoritmus.

5.3 Extrahování dvojic pojmenovaných entit

V této fázi se pomocí značkovače označkují všechny pojmenované entity v korpusu a extrahují se jednotlivé dvojice pojmenovaných entit, včetně jejich značek a kontextu. Dvojice je tvořena dvěma pojmenovanými entitami, které se nacházejí ve stejné větě a zároveň délka kontextu mezi nimi nepřekračuje maximální povolený počet slov. Tato hodnota zároveň definuje maximální možnou délku hledaných parafrází. Dalším požadavkem je, aby se v kontextu mezi nimi nevyskytovala žádná jiná pojmenovaná entita. Kvalita této fáze záleží na kvalitě značkovače, u něhož hraje roli především počet správně rozpoznaných pojmenovaných entit a počet rozpoznávaných kategorií. Jak bylo popsáno v kapitole 3, pro angličtinu byl použit značkovač SuperSense Tagger a pro češtinu byly jako náhrada za neexistující značkovač použity anotační nástroje z PDT 2.0.

V označkovaném textu se za sebou mohou běžně nacházet pojmenované entity, které mají různé značky nebo dokonce obsahují značky dvě (v případě češtiny). Algoritmus je navržen tak, aby extrahoval z textu maximální posloupnost po sobě jdoucích pojmenovaných entit a pak ji prohlásil za jednu víceslovnou pojmenovanou entitu. Pokud se ovšem v posloupnosti nachází různé typy pojmenovaných entit, je těžké vybrat správný typ výsledné entity. Z tohoto důvodu jsou zahozeny dvojice entit, které obsahují alespoň jednu entitu, u níž nelze určit jednoduše typ. Při zpracování anglických dat tato situace nastává zřídkakdy. Horší je to však s češtinou, kde díky nízké kvalitě značkovače tyto situace nastávají velmi často. Abychom pak veškeré pojmenované entity nemuseli degradovat na jednoslovné výrazy, vypočteme si pomocnou tabulku četností všech n-tic pojmenovaných entit v textu. Pro každou vyextrahovanou n-tici entit pak porovnáme četnosti všech jejich podn-tic a nakonec vybereme tu s největší četností. Vše je předvedeno na následujícím příkladě, kde se má vybrat správná podn-tice n-tice `Karel_;Y May_;E_;S Vinnetou_;S_;Y`. Nalezené četnosti jsou uvedeny v tabulce 5.2.

V tomto případě algoritmus vybere správnou hodnotu, kterou je `Vinnetou_;S_;Y`. V mnoha případech ovšem může docházet k chybám, a proto je algoritmus doplněn o některá pravidla či heuristiky, jež mají za úkol nalézt co nejdelší smysluplnou n-tci. Tyto techniky

Podn-tice	Četnost
Karel_;Y May_;E_;S Vinnetou_;S_;Y	1
May_;E_;S Vinnetou_;S_;Y	0
Vinnetou_;S_;Y	144

Tabulka 5.2: Tabulka s četnostmi podn-tic pro určení správné podn-tice pro n-tici Karel May Vinnetou.

jsou použity většinou pouze pro češtinu.

Situace, kdy je pomocí nástrojů z PDT 2.0 přiřazeno jednomu slovu více kategorií, je řešena prioritní tabulkou, kde se vybere vždy kategorie s vyšší prioritou.

Některé kategorie pojmenovaných entit, které značkovače rozpoznají, jsou málo významné nebo podobné některým jiným kategoriím, proto jsou buď ignorovány a nebo sloučeny s jinými kategoriemi. Souhrnný seznam všech použitých kategorií pro češtinu a angličtinu je uveden v tabulce 5.3.

Jazyk	Kategorie
Čeština	PERSON, LOCATION, ORGANIZATION, PRODUCT, CHEMISTRY, MEDICINE, NATURAL_SCIENCE, JUSTICE, TECHNOLOGY, ELECTRONIC, HOBBY, ECONOMY, CULTURE, SPORT, POLICY, ECOLOGY, COLOR
Angličtina	ANIMAL, CONTACT, DISEASE, EVENT, FACILITY, GAME, LOCATION, LANGUAGE, JUSTICE, LOCATION, PERSON, ORGANIZATION, PLANT, PRODUCT, SUBSTANCE, WORK_OF_ART, CARDINAL, MONEY, ORDINAL, QUANTITY DATE, TIME

Tabulka 5.3: Tabulka se všemi kategoriemi pojmenovaných entit pro češtinu a angličtinu.

5.3.1 Formát COUPLES pro uložení dvojic pojmenovaných entit

Protože je extrahování dvojic pojmenovaných entit časově náročné a bylo by nepraktické tento výpočet provádět opakovaně, byl navržen XML formát pro ukládání dvojic pojmenovaných entit společně s dalšími potřebnými informacemi. Tento formát byl nazván *COUPLES* a mírně se liší pro angličtinu a češtinu. V češtině je typ pojmenované entity obsažen již v lemmatu a v angličtině v samostatné značce. Ukázka souboru s tímto formátem pro češtinu je zobrazena na obrázku 5.2.

```
<?xml version="1.0" encoding="ISO-8859-2"?>
<couples>
  <couple count="1">
    <left lemma="Děčín_;G" word="Děčín" />
      <context lemma="doma" word="doma" />
      <context lemma="klopýtnout_:W" word="klopýtl" />
      <context lemma="s-1" word="s" />
```

```
      <right lemma="Ostrava_;G" word="Ostravou" />
    </couple>
  ...
</couples>
```

Obrázek. 5.2: Ukázka formátu *COUPLES* pro češtinu.

Ukázka souboru s tímto formátem pro angličtinu je zobrazena na obrázku 5.3.

```
<?xml version="1.0" encoding="ISO-8859-2"?>
<couples>
  <couple count="3">
      <left lemma="le" word="Le" category="PERSON" />
      <left lemma="pelt" word="Pelt" category="PERSON" />
      <context lemma="throw" word="throws" />
      <context lemma="a" word="a" />
      <right lemma="man" word="man" category="PER_DESC" />
    </couple>
  ...
</couples>
```

Obrázek. 5.3: Ukázka formátu *COUPLES* pro angličtinu.

5.3.2 Experimenty

Z anglických i českých dat se podařilo získat velké množství dvojic pojmenovaných entit. V tabulce 5.4 je uvedeno, kolik dvojic pro jednotlivé jazyky se povedlo získat. Jedná se o dvojice, které mají kontext o maximální délce 5. Jsou zde uvedeny pouze dvojice, které se v textu vyskytly minimálně dvakrát. V tabulce je uveden celkový počet dvojic a také jsou zde uvedeny jednotlivé počty dvojic pro různé délky kontextu. Z tabulky lze vidět relativně srovnatelné množství dvojic pro češtinu i pro angličtinu.

Jazyk	**Délka kontextu**					**Celkem**
	1	2	3	4	5	
Čeština	180200	269088	200218	154454	115707	919667
Angličtina	145389	217220	161760	108262	72164	704795

Tabulka 5.4: Tabulka s počtem extrahovaných dvojic pojmenovaných entit pro jednotlivé jazyky. Rozděleno podle délky kontextu dvojic. Hodnoty v tabulce jsou závislé na nastavení některých parametrů systému.

Ze všech získaných domén byly odfiltrovány takové, které obsahovaly méně než 10 dvojic. Pro angličtinu bylo získáno 301 domén a pro češtinu 83 domén. V tabulce 5.5 je uvedeno 5 domén s největším počtem dvojic pro každý jazyk. Jak jsme předpokládali, pro češtinu bylo nalezeno znatelně méně domén než pro angličtinu. To je způsobeno menším počtem kategorií značkovače pro tento jazyk a také jeho nižší kvalitou.

V Sekineho systému, ze kterého náš systém vychází, bylo z dat získáno 630 tisíc dvojic pojmenovaných entit, které dohromady tvořili 2000 domén.

Jazyk	Doména	Počet dvojic
Čeština	PERSON-PERSON	55567
	LOCATION-LOCATION	29886
	PERSON-LOCATION	29486
	LOCATION-PERSON	25878
	PERSON-ORGANIZATION	10003
Angličtina	PERSON-PERSON	77225
	ORGANIZATION-PERSON	48131
	ORGANIZATION-ORGANIZATION	39424
	PERSON-ORGANIZATION	37665
	PERSON-LOCATION	18015

Tabulka 5.5: Tabulka s ukázkami vybraných domén s největším počtem dvojic pro jednotlivé jazyky. Hodnoty v tabulce jsou závislé na nastavení některých parametrů systému.

5.4 Hledání charakteristických slov pro dvojice pojmenovaných entit

Pro každou doménu se vypočte tabulka charakteristických slov, která ji nejlépe vystihují. Tato slova se hledají v kontextu všech dvojic, kde se pro každé slovo vypočítá jeho skóre podle metriky *tf–idf* popsané například v [19]. Míra *tf–idf* je složena ze dvou dílčích částí *tf* a *idf*. Míra *tf* se snaží maximalizovat skóre pro slova, která se vyskytují v daném dokumentu častěji. Její výpočet odpovídá výpočtu relativní četnosti slova v dokumentu.

$$tf_{w,d} = \frac{n_{w,d}}{\sum_i n_{i,d}}, \tag{5.1}$$

kde w je slovo, d je dokument a $n_{w,d}$ odpovídá četnosti slova w v dokumentu d.

Míra *idf* se snaží penalizovat ohodnocení slov, která se vyskytují ve více dokumentech. Vypočte se podle vzorce

$$idf_w = \log \frac{N}{df_w}, \tag{5.2}$$

kde w je slovo, N je celkový počet dokumentů a df_w odpovídá počtu dokumentů, ve kterých se nachází slovo w.

Výsledná hodnota *tf–idf* je dána součinem obou dílčích měr

$$tf - idf_{w,d} = tf_{w,d} \cdot df_w. \tag{5.3}$$

V našem systému je této metriky využito tak, že dokumentem je chápána doména a slovy dokumentu pak všechna slova jednotlivých kontextů všech dvojic v této doméně. Nevýhodou výpočtu charakteristických slov pomocí metody *tf–idf* je fakt, že na některých pozicích v tabulce se vyskytují slova, která nemají pro danou doménu potřebnou vypovídací hodnotu, např. *a, v, se atd.* pro češtinu nebo *a, and, the atd.* pro angličtinu. Tato slova se umístí do seznamu zakázaných slov a jsou z tabulky odstraněna. V [25] je tato situace řešena pomocí minimální prahové hodnoty *tf–idf*, kterou musí slova v tabulce splňovat. Pokud je jejich *tf–idf* skóre nižší než tento práh, slova se do tabulky nedostanou. Je zřejmé, že v doménách, které obsahují více dvojic, bude pravděpodobně nalezeno větší množství

charakteristických slov. Pro lepší výsledky se výpočet tabulky neprovádí přímo pro slova, ale pro jejich lemmata.

5.4.1 Experimenty

V tabulce 5.6 jsou uvedeny počty nalezených charakteristických slov pro 5 domén s největším počtem dvojic pojmenovaných entit každého jazyka.

Jazyk	Doména	Počet charakteristických slov
Čeština	PERSON-PERSON	12572
	LOCATION-LOCATION	7507
	PERSON-LOCATION	7867
	LOCATION-PERSON	8157
	PERSON-ORGANIZATION	4314
Angličtina	PERSON-PERSON	10937
	ORGANIZATION-PERSON	8344
	ORGANIZATION-ORGANIZATION	7303
	PERSON-ORGANIZATION	7489
	PERSON-LOCATION	4082

Tabulka 5.6: Tabulka s počty charakteristických slov pro vybrané domény obou jazyků. Hodnoty v tabulce jsou závislé na nastavení některých parametrů systému.

Pro srovnání uvedeme, že v Sekineho systému bylo pro doménu *PERSON-PERSON* nalezeno pouhých 618 charakteristických slov a celkem pro všechny domény pak 5184 slov. V našem systému bylo nalezeno výrazně více slov, protože do tabulky umísťujeme všechna slova, která se nenachází v seznamu zakázaných slov, kdežto Sekine pouze slova, která mají *tf–idf* skóre vyšší než je předem stanovená hodnota.

5.5 Vytváření shluků dvojic pojmenovaných entit

V této fázi se provede shlukování frází v každé doméně. Nejprve se každé frázi přiřadí slovo z tabulky charakteristických slov tak, že musí platit, že toto slovo je obsaženo v kontextu dvojice a navíc musí platit, že žádné jiné slovo z jejího kontextu nesmí mít větší skóre v tabulce. Pokud se žádné slovo z kontextu nenachází v tabulce charakteristických slov, pak je fráze zahozena. V dalším kroku spojíme všechny fráze, které mají stejné charakteristické slovo, do jednoho shluku. Na obrázku 5.4 je uveden příklad nalezeného shluku.

5.5.1 Experimenty

V tabulce 5.7 jsou uvedeny počty nalezených shluků pro 5 domén s největším počtem nalezených shluků každého jazyka. Celkový počet domén s alespoň jedním nalezeným shlukem byl pro angličtinu 83 a pro češtinu 18. Minimální požadovaná velikost shluku byla nastavena na hodnotu 10.

Pro srovnání uvádíme, že v Sekineho systému bylo v doméně *COUNTRY-COUNTRY* nalezeno 32 shluků.

Obrázek 5.4: Ukázka shluku získaného v doméně LOCATION-PERSON pro češtinu.

Jazyk	Doména	Počet shluků
Čeština	PERSON-PERSON	288
	PERSON-LOCATION	152
	LOCATION-LOCATION	119
	LOCATION-PERSON	115
	PERSON-ORGANIZATION	36
Angličtina	PERSON-PERSON	341
	ORGANIZATION-PERSON	211
	PERSON-ORGANIZATION	188
	ORGANIZATION-ORGANIZATION	176
	PERSON-LOCATION	88

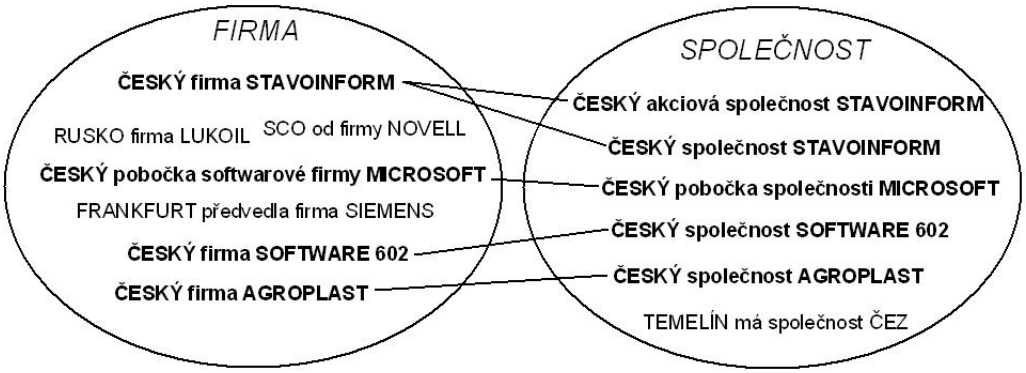
Tabulka 5.7: Tabulka s ukázkami počtu nalezených shluků pro vybrané domény obou jazyků. Hodnoty v tabulce jsou závislé na nastavení některých parametrů systému.

5.6 Hledání vztahů mezi shluky

V posledním kroku algoritmu se provádí vlastní hledání parafrází. V každé doméně se hledají vztahy mezi všemi shluky, na základě instancí dvojic ve shlucích. Pokud se některá instance vyskytuje ve více shlucích, je mezi těmito shluky vytvořen vztah. Při hledání těchto vztahů se mohou hledat shody přímo mezi instancemi nebo mezi lemmaty instancí. Standardně systém provádí hledání vztahů na základě lemmat instancí, čímž docílíme nalezení většího počtu vztahů mezi shluky. Pokud se mezi některými shluky najde více vztahů než je minimální požadovaný počet, pak můžeme mezi těmito shluky nalézt parafráze. Příklad shluků, mezi kterými bylo nalezeno několik vztahů, je na obrázku 5.5 (jedná se o doménu *LOCATION-ORGANIZATION*).

Nyní se zaměříme na to jak lze ze shluků, mezi kterými byly nalezeny vztahy, vytvořit parafráze. Sekine ve své práci [25] považuje za parafráze automaticky všechny fráze v takovýchto shlucích, což by v našem příkladě vedlo k parafrázím

```
LOCATION firma ORGANIZATION,
LOCATION od firmy ORGANIZATION,
LOCATION pobočka softwarové firmy ORGANIZATION,
LOCATION předvedla firma ORGANIZATION,
LOCATION akciová společnost ORGANIZATION,
```



Obrázek 5.5: Diagram znázorňující vztahy mezi shluky, vzniklé na základě stejných lemmat instancí v obou shlucích.

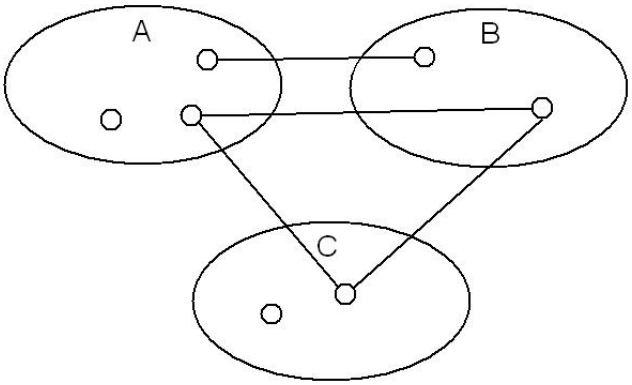
```
LOCATION společnost ORGANIZATION,
LOCATION pobočka společnosti ORGANIZATION,
LOCATION má společnost ORGANIZATION.
```

Jak je vidět, slova *firma* a *společnost* sice jsou parafrázemi, ale všechny fráze obou shluků ne. Pokusíme se tedy tomuto problému vyhnout tím, že mezi parafrázemi zavedeme strukturovanost. Vytvoříme oblasti této množiny frází, z nichž každá oblast bude charakterizována instancemi, pomocí nichž jsme nalezli vztah mezi frázemi této oblasti. Pro fráze, které nemají žádný vztah s jinou frází druhého shluku, oblasti vytvářet nebudeme, proto se ve výsledné množině parafrází neobjeví. Pro uvedenou množinu pak budou vytvořeny oblasti

```
LOCATION firma ORGANIZATION,
LOCATION akciová společnost ORGANIZATION,
LOCATION společnost ORGANIZATION,

LOCATION pobočka softwarové firmy ORGANIZATION,
LOCATION pobočka společnosti ORGANIZATION,

LOCATION firma ORGANIZATION,
LOCATION společnost ORGANIZATION,

LOCATION firma ORGANIZATION,
LOCATION společnost ORGANIZATION.
```

Jak je vidět, popsaný postup může z množiny frází vytvořit několik shodných oblastí. Pokud je některá z nich podmnožinou jiné, odstraníme ji. Výsledná množina pak obsahuje oblasti

```
LOCATION firma ORGANIZATION,
LOCATION akciová společnost ORGANIZATION,
LOCATION společnost ORGANIZATION,

LOCATION pobočka softwarové firmy ORGANIZATION,
```



Obrázek 5.6: Obrázek vysvětlující princip hledání parafrází mezi více shluky.

```
LOCATION pobočka společnosti ORGANIZATION.
```

Touto metodou jsme z množiny frází odstranili některé fráze, které do této množiny nepatřily.

Systém rovněž umí v textu odhalit parafráze, které jsou tvořeny vztahy nejen mezi shluky aktuální domény, ale také vtahy mezi shluky aktuální domény a inverzní domény. To je řešeno tak, že v inverzní doméně se nehledají stejné výskyty instance jako v přímé doméně, ale jsou prohozeny obě části instance. Pokud např. hledáme vztahy mezi shluky pomocí lemmatu instance `<Jelcin_;S, Helsinky_;G>`, budeme v inverzní doméně hledat lemma instance `<Helsinky_;G, Jelcin_;S>`. Pomocí tohoto vylepšení je možné najít některé doposud nenalezené parafráze. Takovéto parafráze budeme v textu označovat symbolem $*$. Příkladem parafrází mezi aktuální a inverzní doménou mohou být fráze

```
PERSON předsedou POLICY,
PERSON jako předseda POLICY,
*POLICY v čele s PERSON.
```

Na rozdíl od systému popsaného v [25] umí tento systém hledat parafráze mezi více než dvěma shluky. Tato situace nastává v případě, že systém nalezne instanci, která se vyskytuje ve více než dvou shlucích. Je tudíž možné nalézt i následující parafráze, které byly vytvořeny ze tří nezávislých shluků. Příkladem mohou být fráze

```
PERSON s manželkou PERSON,
PERSON s chotí PERSON,
PERSON se svou ženou PERSON.
```

Pokud nějaké instance vytváří vztahy mezi více shluky, jsou pak parafráze hledány mezi maximální podmnožinou těchto shluků, která splňuje požadavek minimálního počtu vztahů mezi shluky. Vše ilustruje obrázek 5.6. Pokud bude minimální požadovaný počet vztahů mezi shluky nastaven na hodnotu 1, pak budou hledány parafráze mezi shluky A, B i C. Pokud bude ovšem tato hodnota rovna 2, pak budou parafráze hledány pouze mezi shluky A a B.

5.6.1 Formát PARAPHRASES pro uložení nalezených parafrází

Pro ukládání nalezených parafrází byl navržen XML formát nazvaný *PARAPHRASES*. Tento formát umožňuje uložení parafrází v jejich strukturované podobě spolu s instancemi, pomocí kterých byly nalezeny. Formát rovněž počítá s dalším použitím při vyhodnocování parafrází, a proto do něj lze uložit i informace o správnosti jednotlivých frází či vztahů. Ukázka souboru s tímto formátem je uvedena na obrázku 5.7.

```
<?xml version="1.0" encoding="ISO-8859-2"?>
<paraphrases>
  <domain id="LOCATION - CARDINAL">
    <paraphrase>
      <link>
        <cluster id="split" correct="true" />
        <cluster id="divide" correct="true" />
      </link>
      <sublink>
        <instance left="village" right="two" />
        <pattern value="LOCATION is split in LOCATION"
           correct="true" />
        <separator />
        <pattern value="LOCATION is divided into LOCATION
           " correct="true" />
        <pattern value="LOCATION divided into" correct="
           true" />
        <separator />
      </sublink>
      ...
    </paraphrase>
    ...
  </domain>
  ...
<paraphrases>
```

Obrázek. 5.7: Ukázka formátu *PARAPHRASES*.

5.6.2 Experimenty

V tabulce 5.8 jsou uvedeny počty nalezených parafrází pro 5 domén s největším počtem nalezených parafrází každého jazyka. Celkový počet nalezených parafrází pro dané nastavení parametrů byl pro angličtinu 90 a pro češtinu 182. Počet domén, ve kterých byly nalezeny parafráze, byl pro angličtinu i pro češtinu 15.

Jazyk	Doména	Počet parafrází
Čeština	LOCATION-PERSON	74
	PERSON-PERSON	33
	LOCATION-LOCATION	28
	PERSON-LOCATION	24
	PERSON-POLICY	24
Angličtina	CARDINAL-PERSON	117
	CARDINAL-ORGANIZATION	25
	ORDINAL-DATE	11
	ORGANIZATION-DATE	6
	DATE-ORDINAL	4

Tabulka 5.8: Tabulka s ukázkami počtu nalezených parafrází pro vybrané domény obou jazyků. Hodnoty v tabulce jsou závislé na nastavení některých parametrů systému.

Kapitola 6

Vyhodnocení a výsledky systému

V této kapitole bude uveden přehled metod pro vyhodnocování systémů, které pracují s parafrázemi. Bude zde popsána metoda použitá při vyhodnocování našeho systému a budou zde uvedeny i dosažené výsledky. Součástí této kapitoly je rovněž citlivostní analýza našeho systému a lokalizace jeho slabých míst. Na závěr této kapitoly pak zmíníme několik technických aspektů našeho systému.

6.1 Metody vyhodnocování

Ačkoliv je použití parafrází téměř nutností pro správnou funkci mnohých algoritmů v oblasti zpracování přirozeného jazyka, doposud nebylo uspokojivě vyřešeno vyhodnocování systémů, které s parafrázemi pracují. Každý tvůrce systému má většinou své vlastní metody vyhodnocování a jen velmi těžko lze systém srovnávat s ostatními. Ve většině případů se vyhodnocování provádí pomocí lidských hodnotitelů, což vede k velké časové náročnosti a nižší spolehlivosti tohoto procesu. Při vyhodnocování systémů pro získávání parafrází se kvůli neexistenci kvalitních bází znalostí či testovacích sad bez tohoto způsobu téměř neobejdeme. Oproti tomu na poli vyhodnocování klasifikátorů parafrází již snahy o zautomatizování tohoto procesu najít můžeme. To je do značné míry způsobeno jednodušší tvorbou testovacích datových sad než u systému pro získávání parafrází. Asi nejznámějším pokusem o zautomatizování vyhodnocování klasifikátorů parafrází je vytvoření anotovaného parafrázového korpusu Microsoft Research Paraphrase Corpus, popsaného v [12], nebo korpusu budovaného v rámci semináře PASCAL Challenges Workshop on Recognising Textual Entailment.

6.1.1 Systémy pro získávání parafrází

Jak již bylo zmíněno, vyhodnocování systému pro získávání parafrází probíhá zpravidla ručně pomocí lidských hodnotitelů. Tento způsob je časově i finančně náročný, v závislosti na hodnotiteli také značně subjektivní. Vyhodnocování systému provádí často více hodnotitelů, kteří se nemusí shodovat ve svých odpovědích. Je proto dobré nějakým způsobem umět vyhodnocovat i shodu mezi hodnotiteli. Tato situace se řeší pomocí *kappa indexu*. Informace o *kappa indexu* pocházejí z [19]. Mějme výsledky vyhodnocování systému dvěma hodnotiteli A a B, které jsou znázorněny v tabulce 6.1.
Pro hodnoty g_1, g_2, f_1 a f_2 z tabulky pak platí vztahy

		A		
		ano	**ne**	**celkem**
B	**ano**	a	b	g_1
	ne	c	d	g_2
	celkem	f_1	f_2	N

Tabulka 6.1: Tabulka shody při vyhodnocování systému hodnoteli A a B.

$$g_1 = a + b, \tag{6.1}$$
$$g_2 = c + d, \tag{6.2}$$
$$f_1 = a + c, \tag{6.3}$$
$$f_2 = b + d, \tag{6.4}$$
$$N = a + b + c + d = f_1 + f_2 = g_1 + g_2, \tag{6.5}$$

kde a je počet frází, které oba hodnotitelé považují za parafráze, b je počet frází, které hodnotitel B považuje za parafráze, ale hodnotitel A nikoliv, c je počet frází, která hodnotitel A považuje za parafráze, ale hodnotitel B nikoliv a d je počet frází, které žádný z hodnotitelů nepovažuje za parafráze.

Celkový poměr shody obou hodnotitelů je dán vztahem

$$P(A) = \frac{a + d}{N}. \tag{6.6}$$

Celkový očekávaný poměr jejich shody vztahem

$$P(E) = \left(\frac{f_1 + g_1}{2 \cdot N}\right)^2 + \left(\frac{f_2 + g_2}{2 \cdot N}\right)^2. \tag{6.7}$$

Hodnota *kappa indexu* je pak dána vztahem

$$kappa = \frac{P(A) - P(E)}{1 - P(E)}. \tag{6.8}$$

Tato hodnota nám slouží jako míra shody dvou hodnotitelů. Pokud je mezi oběma hodnotiteli přesná shoda, pak je hodnota *kappa indexu* rovna 1, pokud je mezi nimi čistě náhodná shoda, je *kappa index* 0 a pokud je míra shody mezi oběma hodnotiteli horší než náhodná, je hodnota *kappa indexu* záporná. V tabulce 6.2 jsou popsány významy intervalů *kappa indexu*.

Hodnota *kappa indexu*	**Význam**
$> 0,8$	dobrá shoda
$0,67 - 0,8$	přiměřená shoda
$0,67 >$	data poskytují pochybnou bázi pro vyhodnocování

Tabulka 6.2: Tabulka s hodnotami *kappa indexu* a jejich interpretací.

Pokud je potřeba určit míru shody pro více než dva hodnotitele, provádí se to většinou pomocí aritmetického průměru *kappa indexů* všech párů hodnotitelů.

Je nutné rovněž zvážit, zda hodnotitelům v průběhu procesu vyhodnocování ukázat parafráze v kontextu celé věty, nebo kontext zatajit. V mnoha případech může tato skutečnost změnit názor hodnotitele. Aby bylo dosaženo co největší spolehlivosti procesu, bývají hodnotitelé často vybíráni z řad zkušených rodilých mluvčích, ne-li přímo lingvistů.

Při vyhodnocování systémů je potřeba měřit jejich přesnost P (*precision*), která udává, kolik nalezených skupin frází je ve skutečnosti parafrázemi, a také pokrytí R (*recall*), které udává, kolik parafrází z celkového počtu všech parafrází v textu bylo nalezeno. V případě vyhodnocování přesnosti stačí pouze klasifikovat, jestli systémem nalezené fráze jsou parafrázemi či nikoliv. V případě měření pokrytí je potřeba projít celá testovací data a vyhledat v nich všechny parafráze, což může být časově velmi náročné a proto se od toho často upouští. Výpočet přesnosti a pokrytí probíhá podle vzorců

$$P = \frac{t_p}{t_p + f_p}, \tag{6.9}$$

$$R = \frac{t_p}{t_p + f_n}, \tag{6.10}$$

kde t_p označuje počet nalezených skupin frází, které jsou parafrázemi, f_p počet nalezených skupin frází, které nejsou parafrázemi a f_n počet skupin parafrází v datech, které nebyly nalezeny.

Úspěšnost systémů bývá velmi závislá na použitých datech a algoritmech. Např. v [9], kde byl použit klasifikátor Support Vector Machines a data z novinových článků, bylo dosaženo maximální přesnosti 87,42% a pokrytí 87,66%.

Zvláštní zmínku si zaslouží také způsob vyhodnocování použitý v [7], kde není účelem vyhodnocovat vlastní parafráze, ale parafráze jsou zde brány jako orientované relace mezi textovými částmi, u kterých je nutné určit i správný směr. Vyhodnocení tohoto problému je silně závislé na kontextu. Zde bylo dosaženo přesnosti 44,15%.

Někdy se lze setkat se systémy, u nichž nejsou vyhodnocovány přímo parafráze, ale některé vnitřní stavy systému jako např. kvalita klasifikátoru příznaků (viz. [24]) či kvalita zarovnání (viz. [9], [4]). Do této kategorie také patří vyhodnocování pomocí systémů využívajících parafráze. Příkladem takového vyhodnocení může být u systému pro získávání informací, který je popsán v [22]. Nejprve se ohodnotí řešení nalezené systémem bez použití parafrází a následně se vyhodnotí řešení systému s použitím parafrází. Vyhodnocování bývá zpravidla ruční a hodnotitel musí umět vyhodnotit a změřit kvalitu nalezeného řešení. Tento systém dosahoval úspěšnosti 52,7% bez použití parafrází a 71,73% s použitím parafrází. Dalším příkladem může být vyhodnocování pomocí systému pro generování parafrází. V případě nově vygenerované parafráze může být vygenerovaná věta považována za jeden z následujících případů:

- Úplná parafráze bez informací navíc.
- Úplná parafráze s informací navíc.
- Částečná parafráze bez informací navíc.
- Částečná parafráze s informací navíc.

Hodnotitelé pak mohou klasifikovat vygenerované parafráze do jedné ze zmíněných kategorií a na základě toho určit skóre systému.

6.1.2 Systémy pro klasifikaci parafrází

Klasifikátory parafrází bývají vyhodnocovány zpravidla automaticky pomocí ručně vytvořených anotovaných korpusů. Kvalita a dostatek takovýchto korpusů ovšem není v dnešní době uspokojivá. Mezi jeden z mála použitelných korpusů parafrází patří Microsoft Research Paraphrase Corpus nebo korpus vzniklý v rámci semináře PASCAL Challenges Workshop on Recognising Textual Entailment konaného v roce 2005. Úspěšnosti klasifikátorů jsou opět silně závislé na použitých datech a algoritmech. Jako příklad lze uvést systém uvedený v [24], který byl trénován a testován na Microsoft Research Paraphrase Corpus a dosahoval celkové přesnosti 72,5% a pokrytí 93,4%.

Microsoft Research Paraphrase Corpus

Microsoft Research Paraphrase Corpus, který je volně ke stažení[1], byl vytvořen autory Williamem B. Dolanem and Chrisem Brockettem a je popsán v jejich práci [12]. Snahou bylo vytvořit korpus, který by mohl sloužit k trénování a testování klasifikátorů parafrází.

Korpus se skládá z 5801 dvojic vět, které jsou anotovány lidskými anotátory. U každé dvojice je vždy označeno, zda se jedná o parafráze, či nikoliv. Korpus obsahuje 3900 dvojic (67%) označených jako parafráze a 1901 dvojic (33%), které nejsou parafrázemi. Korpus byl vybudován z porovnatelného korpusu novinových článků použitím několika heuristik, jejichž bližší podrobnosti lze nalézt v [12]. Při budování korpusu bylo na věty kladeno několik požadavků, které měly především za úkol zajistit diverzitu nalezených vět a zmenšit prohledávaný prostor pozdějšího ručního anotování. Mezi tyto požadavky patří např. délka vět v určitém rozsahu, minimální počet sdílených slov ve větách, délky vět v určitém poměru nebo definovaná lexikální vzdálenost slov ve větách.

Tento korpus byl zkoumán několika výzkumníky, kteří zjistili, že má poměrně velký lexikální překryv (cca 70%); hlavně u dvojic vět, které jsou označeny jako parafráze. Proto mají uvedené dvojice malou lexikální diverzitu. Autoři sami přiznávají, že distribuce parafrází uvnitř korpusu neodráží reálnou distribuci a je nevhodné použít korpus k trénování klasifikátorů. Korpus je proto využíván spíše k testování systémů, které klasifikují parafráze. Ačkoliv autoři slibují nápravu a plánují korpus rozšířit, ke stažení je dostupná pouze původní verze 1.0, u které byly zjištěny tyto nedostatky.

Podobnou práci můžeme nalézt v [13], kde se autoři pokoušeli vytvořit podobný korpus parafrází v japonštině. Korpus je rozdělen do několika oddílů, které reprezentují odlišné třídy parafrází a snaží se postihnout jejich pokrytí v reálných textech. Korpus byl rovněž vytvořen z novinových článků a tvoří současný standard pro japonštinu.

PASCAL Challenges Workshop on Recognising Textual Entailment

Tento korpus je popsán v [11] a je volně stažitelný na internetu[2]. Jedná se o korpus, který obsahuje dvojice úryvků textu, z nichž jeden je brán jako text T a druhý jako hypotéza H (viz 2.1.3). T představuje typicky jednu nebo dvě věty, zatímco H je pouze malým útržkem věty. Dvojice je vždy ručně anotována informacemi o tom, jestli tvoří orientovanou relaci či nikoliv a také obsahuje informaci o kategorii oblasti využití daného příkladu. Kategoriemi, kterých je celkem sedm, jsou získávání informací (IR), porovnatelné dokumenty (CD), porozumění textu (RC), odpovídání na otázky (QA), extrakce informací (IE), strojový překlad

[1] http://research.microsoft.com/en-us/downloads/607D14D9-20CD-47E3-85BC-A2F65CD28042/default.aspx

[2] http://www.pascal-network.org/Challenges/RTE/Datasets/

(MT) a získávání parafrází (PP). Uvnitř každé skupiny se vyskytují jak kladné, tak záporné příklady. Všechny páry byly získány ručně a jsou rozděleny na trénovací část a testovací část. Trénovací část obsahuje 567 dvojic a testovací část 800 dvojic. Ačkoliv je distribuce příkladů odlišná od distribuce v reálných textech, tvůrci se snažili vyhnout korelaci mezi lexikálním překryvem a zařazením do třídy. Příklady dvojic v datové sadě jsou uvedeny v tabulce 6.3.

ID	TEXT	HYPOTHESIS	TASK	ENTAILMENT
1	iTunes software has seen strong sales in Europe.	Strong sales for iTunes in Europe.	IR	True
2	Cavern Club sessions paid the Beatles L15 evenings and L5 lunchtime.	The Beatles perform at Cavern Club at lunchtime.	IR	True
3	American Airlines began laying off hundreds of flight attendants on Tuesday, after a federal judge turned aside a union's bid to block the job losses.	American Airlines will recall hundreds of flight attendants as it steps up the number of flights it operates.	PP	False
4	Paul Bremer, the top U.S. civilian administrator in Iraq, and Iraq's new president, Ghazi al-Yawar, visited the northern Iraqi city of Kirkuk.	Ghazi al-Yawar is Iraq's new president.	RC	True
5	Baghdad had announced that it will stop cooperating with UNSCOM completely but indicated that it will not ask for their departure.	Baghdad announced the complete halt in their cooperation with UNSCOM, and said also, that it will ask them to leave.	MT	False

Tabulka 6.3: Příklady dvojic v datové sadě k semináři PASCAL Challenges Workshop on Recognising Textual Entailment.

6.1.3 Pracovní semináře týkající se parafrází

Mezinárodní organizace ACL[3] pořádala celkem již tři semináře s názvem International Workshop on Paraphrasing (IWP) týkající se parafrází. Semináře se konaly v roce 2001 [1], 2003 [2] a 2005 [3] a jejich cílem bylo motivovat výzkumné pracovníky k vylepšení technik systému, které se zabývají parafrázemi.

[3] Association for Computational Linguistics

Mezi témata, která se na seminářích probírala, patří např. vlastní podstata parafrází, rozdíly mezi parafrázemi a shrnutím, parafráze na úrovni slov, vět i celých článků, algoritmy pro generování, rozpoznávání a výběr parafrází, aplikace parafrází v reálných systémech, tvorba báze znalostí pro automatickou práci s parafrázemi, automatického získávání parafrází, počítačového modelování lingvistické teorie parafrází, vyhodnocování algoritmů a zdrojů parafrází a mnoho dalších.

S mnohými poznatky z těchto seminářů se čtenář může setkat i v této diplomové práci.

6.2 Evaluace systému a diskuze výsledků

Z předchozího textu vyplynulo, že vyhodnocení systému získavajícího parafráze z textu není jednoduché. Náš systém bude stejně jako většina podobných vyhodnocen pomocí lidských hodnotitelů. Jelikož vstupní data, ze kterých se systém snažil nalézt parafráze, dosahují řádově velikosti GB, nebude u systému prováděno vyhodnocení pokrytí, ale pouze přesnosti nalezených parafrází. V podobném systému, který je popsán v [25], je prováděno rovněž pouze měření přesnosti a to dvojí – přesnosti shluků a přesnosti vztahů mezi shluky. Přesnost shluků lze v tomto případě chápat jako čistotu nalezených parafrází a přesnost vztahů mezi shluky pak jako správnost parafrází. U našeho systému se zaměříme rovněž na obě tyto přesnosti. Nejprve je ale nutné definovat přesná kritéria pro vyhodnocování obou těchto přesností. Při vyhodnocování nebude hodnotitelům ukazován celý kontext parafrází, protože tato informace se ztrácí již při extrahování dvojic pojmenovaných entit.

6.2.1 Kritéria vyhodnocování

V [25] jsou dvě fráze považovány za parafráze, pokud je můžeme použít pro vyjádření stejného vztahu mezi dvěma entitami v oblasti extrakce informací. Našim cílem není hledat všechny fráze, které vyjadřují stejné nebo podobné relace mezi entitami, ale hledat parafráze. Při vyhodnocování systému nám půjde zejména o to, zda lze dané fráze v textu zaměnit bez toho, aniž bysme změnili význam textu a nebo se dopustili gramatické chyby.

Ve shlucích se často nacházejí fráze, které obsahují stejné charakteristické slovo, ale vyjadřují mezi entitami odlišné relace. Tyto fráze mohou být v některých případech považovány za parafráze nebo alespoň jejich části, ale k tomuto zjištění je v drtivé většině případů potřeba znát celé kontexty vět, z nichž byly parafráze vytvořeny. Takové fráze nebudeme při vyhodnocování považovat za parafráze. Příkladem takových frází mohou být fráze

```
LOCATION při vstupu do ORGANIZATION,
LOCATION ke vstupu do ORGANIZATION,
LOCATION po vstupu do ORGANIZATION,
LOCATION o vstup do ORGANIZATION,
LOCATION na vstup ORGANIZATION.
```

Za parafráze nebudeme rovněž považovat fráze, které obsahují jako charakteristické slovo sloveso a liší se pouze jeho časem. Takové fráze mohou vypadat např. takto

```
ORGANIZATION koupila PRODUCT,
ORGANIZATION bude kupovat PRODUCT.
```

Rovněž se často setkáme s frázemi, které by mohly být považovány za parafráze, ale liší se použitou osobou, což nám neumožňuje vzájemnou záměnu obou frází v textu bez toho,

aniž bychom se dopustili gramatické chyby. Tento případ se týká hlavně češtiny. Příslušné fráze, demonstrující tento problém, jsou

```
PERSON to včera sdělila ORGANIZATION,
PERSON to včera sdělil ORGANIZATION.
```

Velmi často se také ve skupině parafrází vyskytují fráze, které sice vyjadřují stejný vztah mezi entitami, ale v jiném pádě. Takovéto fráze tedy nemůžeme ve větě korektně zaměnit, proto je za parafráze nebudeme považovat. Příkladem takových frází mohou být fráze

```
PERSON a jeho manželka PERSON,
PERSON a jeho manželky PERSON,
PERSON a jeho manželku PERSON.
```

Dalším případem, který je potřeba zvážit při vyhodnocování, je jak vyhodnocovat orientované relace mezi parafrázemi. Při vyhodnocování těchto relací v našem systému musí platit, aby obecnější fráze šla v textu dosadit za specifičtější. Příkladem takovýchto parafrází mohou být fráze

```
LOCATION pobočka americké softwarové firmy ORGANIZATION,
LOCATION pobočka společnosti ORGANIZATION.
```

Je zřejmé, že všude, kde se vyskytuje první fráze, můžeme dosadit i druhou, naopak to ovšem neplatí. Jedná se o typický příklad orientované relace mezi parafrázemi. Takovéto případy budeme považovat za parafráze, i když je možné je použít jen jednostranně.

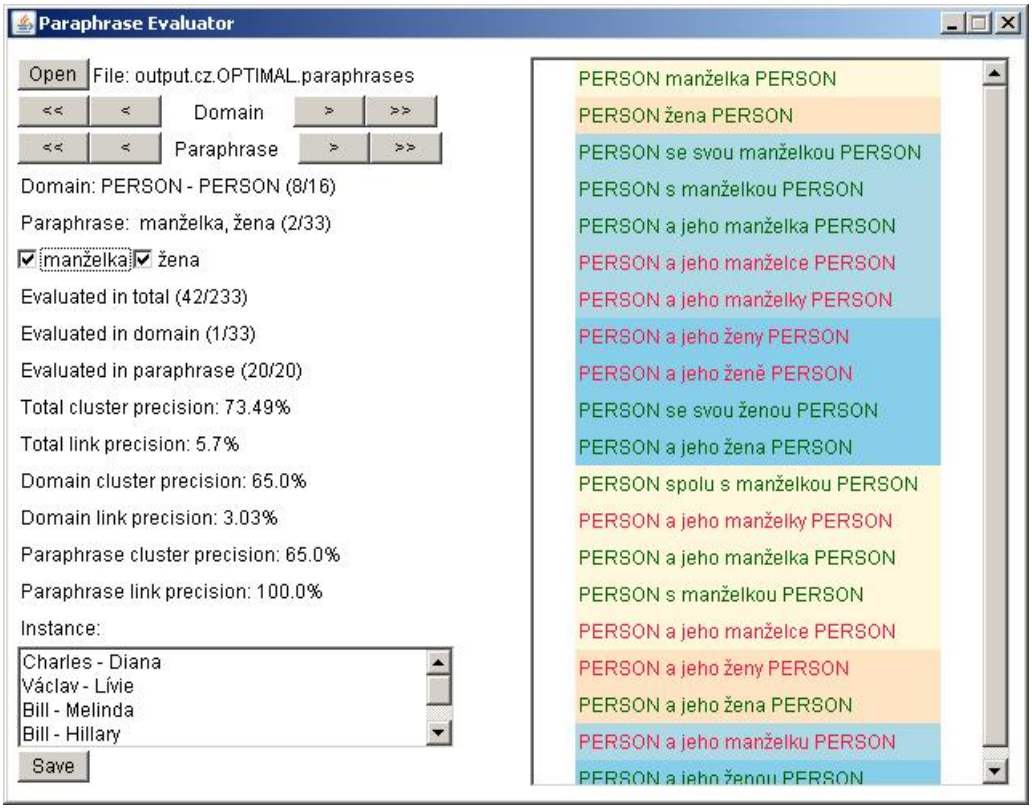
V [25] je brána přesnost shluku jako poměr počtu skutečných parafrází k celkovému počtu frází ve shluku. Vztah mezi dvěma shluky je považován za správný, pokud se většina frází ze shluku dá použít ve stejném kontextu jako většina frází z druhého shluku. Jelikož jsou parafráze nalezené naším systémem mírně odlišné od parafrází v tomto systému, je potřeba tato kritéria upravit. Při výpočtu přesnosti shluků je potřeba brát v potaz fakt, že náš systém hledá strukturované parafráze. Parafráze je tedy tvořena několika oblastmi. Celková přesnost shluku je pak brána jako aritmetický průměr přesností všech oblastí parafrází. Situace je mírně odlišná od Sekineho systému i v případě výpočtu přesnosti vztahů mezi shluky. Parafráze v našem systému totiž mohou být tvořeny vztahy mezi více než dvěma shluky. Pokud tomu tak je, pak je přesnost dána poměrem počtu správných shluků k celkovému počtu shluků. Pokud pro danou skupinu parafrází vyjde přesnost vzahů mezi shluky rovna 0, pak je zřejmé, že množina frází ve skutečnosti nejsou parafráze, a proto v tomto případě nebudeme vyhodnocovat přesnost shluků.

V tabulce 6.4 je uvedeno srovnání našeho a Sekineho systému z pohledu toho, jaký typ frází systém považuje za parafráze. Tabulka nepokrývá všechny možné typy frází.

6.2.2 Grafické rozhraní pro vyhodnocování

Jelikož je ruční vyhodnocování velmi náročné, bylo do systému implementováno grafické rozhraní, jehož účelem je vyhodnocování co nejvíce usnadnit. Hodnotitel se tak nemusí zaměřovat na výpočty jednotlivých přesností, ale pouze označuje korektnost jednotlivých frází ve shlucích či korektnost vztahů mezi shluky. Systém automaticky provádí výpočet všech přesností. Ukázka grafického rozhraní je zobrazena na obrázku 6.1.

Pomocí tohoto rozhraní lze provést vyhodnocení parafrází, které jsou uloženy ve formátu *PARAPHRASES*. Vyhodnocené parafráze lze v tomto formátu rovněž uložit.



Obrázek 6.1: Ukázka grafického rozhraní programu pro vyhodnocování výsledků.

Popis frází	Náš systém	Sekineho systém
fráze se stejným klíčovým slovem, které vyjadřují odlišnou relaci	ne	někdy
fráze vyjadřující stejnou relaci, ale jiný čas	ne	ano
fráze vyjadřující stejnou relaci, ale jinou osobu	ne	ano
fráze vyjadřující stejnou relaci, ale jiný pád	ne	někdy
orientované relace mezi parafrázemi	ano	někdy

Tabulka 6.4: Srovnání našeho systému a Sekineho systému popsaného v [25] z pohledu toho, co který systém považuje za parafráze.

6.2.3 Citlivostní analýza a výsledky systému

V této části se zaměříme na vyhodnocování vlivu některých parametrů systému na přesnost a počet nalezených parafrází. Tato analýza nám pomůže nalézt optimální volbu těchto parametrů, která zaručí maximální přesnost nalezených parafrází, ale také si na základě ní můžeme udělat představu, jak který parametr ovlivní výsledky systému. Jelikož jsou zdroje dat pro češtinu a angličtinu značně odlišné, bude potřeba tuto analýzu provést samostatně pro oba dva jazyky. Tato analýza zahrnuje mnoho vyhodnocování výstupů systému pro různá nastavení parametrů. Kvůli časové náročnosti bude provedeno vyhodnocování pouze pro vybrané domény pro každý jazyk. Pro češtinu byly vybrány domény *PERSON-LOCATION* a *PERSON-PERSON* a pro angličtinu domény *CARDINAL-ORGANIZATION* a *ORDINAL-DATE*. Měření vlivu určitého parametru je prováděno při konstantních hodnotách všech ostatních parametrů.

Počet vztahů mezi shluky

Jeden z nejdůležitějších parametrů v systému je parametr, který určuje minimální hodnotu počtu vztahů mezi shluky, aby byly fráze ve shlucích považovány za parafráze. Tato hodnota výrazně ovlivňuje počet nalezených parafrází a také jejich přesnost. Hodnoty naměřené při testování tohoto parametru jsou uvedeny v tabulkách 6.5 a 6.6. V tabulce pro češtinu nejsou změřeny hodnoty pro nastavení tohoto parametru na hodnotu *1*, protože v daných doménách bylo nalezeno extrémně velké množství parafrází.

Z tabulky tohoto parametru pro češtinu je vidět, že mezi počtem nalezených parafrází a jejich přesností, v závislosti na tomto parametru, je nepřímá úměrnost. Lze předpokládat, že přesnost parafrází pro nastavení tohoto parametru na hodnotu *1* by v případě, že by byla vyhodnocena, byla nejmenší ze všech. V [25] je tento parametr nastaven na hodnotu *2*. Tuto hodnotu lze v našem systému rovněž použít, upřednostníme ale přesnost nalezených parafrází, a proto budeme brát jako optimální hodnotu *3*.

Z tabulky 6.6 je vidět, že stejně jako pro češtinu, tak i pro angličtinu lze dosáhnout lepší přesnosti na úkor počtu nalezených parafrází. I zde je vhodné nastavit tento parametr na hodnotu *3*.

hodnota parametru	Čeština					
	PERSON-LOCATION			PERSON-PERSON		
	P_V	P_S	N	P_V	P_S	N
1	-	-	653	-	-	641
2	25,29%	81,45%	87	39,20%	87,60%	82
3	37,50%	84,90%	24	51,52%	84,71%	33
4	58,33%	80,33%	12	45,00%	83,11%	20
5	71,43%	74,84%	7	53,33%	80,21%	15

Tabulka 6.5: Vliv minimálního počtu vztahů mezi shluky na přesnost parafrází pro vybrané domény pro češtinu.

hodnota parametru	Angličtina					
	CARDINAL-ORGANIZATION			ORDINAL-DATE		
	P_V	P_S	N	P_V	P_S	N
1	37,86%	100,00%	140	23,73%	83,93%	59
2	53,06%	100,00%	49	33,33%	81,25%	24
3	68,00%	100,00%	25	36,36%	75,75%	11
4	75,00%	100,00%	12	40,00%	50,00%	5
5	83,33%	100,00%	6	25,00%	50,00%	4

Tabulka 6.6: Vliv minimálního počtu vztahů mezi shluky na přesnost parafrází pro vybrané domény pro angličtinu.

Délka parafrází

Dalším parametrem, jehož vliv byl testován, je maximální délka nalezených parafrází. Tento parametr způsobí to, že veškeré dvojice, které mají délku kontextu větší než je hodnota tohoto parametru, jsou ještě před vlastním hledáním parafrází zahozeny. Hodnoty přesností parafrází pro různé hodnoty tohoto parametru lze nalézt v tabulkách 6.7 a 6.8.

hodnota parametru	Čeština					
	PERSON-LOCATION			PERSON-PERSON		
	P_V	P_S	N	P_V	P_S	N
1	100,0%	66,67%	1	27,78%	100,00%	18
2	50,00%	80,44%	12	48,00%	94,44%	25
3	31,82%	79,39%	22	51,72%	84,05%	29
4	36,36%	55,95%	22	53,13%	85,82%	32
5	37,50%	84,90%	24	51,52%	84,71%	33

Tabulka 6.7: Vliv délky parafrází na jejich přesnost pro vybrané domény pro češtinu.

Z obou uvedených tabulek je vidět, že nastavení tohoto parametru na hodnotu menší než *3* vede k výraznému poklesu počtu nalezených parafrází. Pro hodnoty *3*, *4* a *5* se pak přesnosti parafrází nijak výrazně neliší, proto je vhodné tento parametr pro oba jazyky nastavit na hodnotu *5*, což zajistí maximální počet nalezených parafrází.

hodnota parametru	Angličtina					
	CARDINAL-ORGANIZATION			ORDINAL-DATE		
	P_V	P_S	N	P_V	P_S	N
1	0,00%	0,00%	0	0,00%	0,00%	0
2	75,00%	100,00%	4	16,67%	100,00%	6
3	66,67%	100,00%	18	37,50%	83,33%	8
4	68,00%	100,00%	25	44,44%	81,25%	9
5	68,00%	100,00%	25	36,36%	75,75%	11

Tabulka 6.8: Vliv délky parafrází na jejich přesnost pro vybrané domény pro angličtinu.

Počet výskytů dvojice pojmenovaných entit v datech

Dalším testovaným parametrem je minimální počet výskytů dvojice pojmenovaných entit v datech, aby tato dvojice nebyla zamítnuta. Při výpočtu se počítá pouze s dvojicemi, které splňují minimální stanovaný počet výskytů. Hodnoty přesností pro testování vlivu tohoto parametru jsou uvedeny v tabulkách 6.9 a 6.10. Při nastavení parametru na hodnotu *1* se výpočet kvůli paměťové náročnosti nepovedlo dokončit.

hodnota parametru	Čeština					
	PERSON-LOCATION			PERSON-PERSON		
	P_V	P_S	N	P_V	P_S	N
1	-	-	-	-	-	-
2	37,50%	84,90%	24	51,52%	84,71%	33
3	50,00%	100,00%	2	14,29%	100,00%	7
4	0,00%	0,00%	0	0,00%	0,00%	2

Tabulka 6.9: Vliv minimálního počtu výskytů dvojic pojmenovaných entit na přesnost parafrází pro vybrané domény pro češtinu.

hodnota parametru	Angličtina					
	CARDINAL-ORGANIZATION			ORDINAL-DATE		
	P_V	P_S	N	P_V	P_S	N
1	52,50%	100,00%	40	44,44%	81,25%	9
2	68,00%	100,00%	25	36,36%	75,75%	11
3	56,25%	100,00%	8	0,00%	0,00%	0

Tabulka 6.10: Vliv minimálního počtu výskytů dvojic pojmenovaných entit na přesnost parafrází pro vybrané domény pro angličtinu.

Dá se předpokládat, že čím vyšší hodnotu tohoto parametru nastavíme, tím vyšší bude dosažená přesnost a zároveň dojde ke snížení počtu nalezených parafrází. Jelikož při měření vlivu tohoto parametru pro hodnoty vyšší než *2* nebyly nalezeny téměř žádné parafráze, lze doporučit pro oba jazyky nastavit tento parametr na hodnotu *2*.

Instance vs. lemmata instancí

Poslední testovaný parametr určuje, zda se vztahy mezi shluky budou hledat na základě instancí nebo lemmat instancí. Vliv tohoto parametru na přesnost systému lze vyčíst z tabulek 6.11 a 6.12. Hodnota parametru *LEMMA* udává, že budou použita lemmata instance a hodnota *WORD* udává, že budou použity instance.

hodnota parametru	**Čeština**					
	PERSON-LOCATION			**PERSON-PERSON**		
	P_V	P_S	N	P_V	P_S	N
WORD	50,00%	88,89%	6	54,55%	93,85%	22
LEMMA	37,50%	84,90%	24	51,52%	84,71%	33

Tabulka 6.11: Vliv instancí a lemmat instancí na přesnost parafrází pro vybrané domény pro češtinu.

hodnota parametru	**Angličtina**					
	CARDINAL-ORGANIZATION			**ORDINAL-DATE**		
	P_V	P_S	N	P_V	P_S	N
WORD	80,95%	100,00%	21	14,29%	100,00%	7
LEMMA	68,00%	100,00%	25	36,36%	75,75%	11

Tabulka 6.12: Vliv instancí a lemmat instancí na přesnost parafrází pro vybrané domény pro angličtinu.

Z naměřených výsledků lze vidět, že nastavením parametru na hodnotu *WORD* dosahuje systém většinou lepší přesnosti než při hodnotě *LEMMA*. Tento jev je více patrný spíše u češtiny, která má mnohem komplikovanější morfologii. Ačkoliv je dosaženo lepší přesnosti, opět to je na úkor počtu nalezených parafrází. Pro oba jazyky lze tedy doporučit tento parametr nastavit na hodnotu *LEMMA*, aby bylo nalezeno dostatečné množství parafrází.

Optimální výsledky systému

Nejlepších výsledků při hledání parafrází lze docílit, pokud využijeme provedené citlivostní analýzy a nastavíme hodnoty všech parametrů na optimální hodnoty. Na závěr uvádíme počty parafrází a jejich přesnosti pro *5* domén s největším počtem nalezených parafrází pro každý jazyk. Při nalezení těchto parafrází bylo použito optimální nastavení parametrů.

Z tabulky 6.13 lze vidět, že systémem nalezené parafráze v angličtině mají obě přesnosti větší než v češtině. Sekine pro svůj systém uvádí dosažené přesnosti shluků od 65% do 99% a přesnosti vztahů mezi shluky pak od 73% do 86%. Pokud vezmeme v úvahu srovnání obou systémů z hlediska toho, co který považuje za parafráze (viz tabulka 6.4), pak je zřejmé, že mnoho frází, které jsou v Sekineho práci považovány za parafráze, v našem systému považovány za parafráze nejsou. Kritéria pro rozhodování o tom, zda se jedná či nejedná o parafráze, jsou u obou systémů stanovena jinak. Lze říci, že v našem systému jsou tato kritéria mnohdy přísnější. Tato kritéria se týkají zejména vyhodnocování přesnosti uvnitř shluků a neměla by nijak výrazně ovlivnit přesnost vztahů mezi shluky. I přesto jsou ovšem

Jazyk	Doména	P_V	P_S	N
Čeština	LOCATION-PERSON	40,54%	82,00%	74
	PERSON-PERSON	51,52%	84,71%	33
	LOCATION-LOCATION	14,29%	91,67%	28
	PERSON-LOCATION	37,50%	84,90%	24
	PERSON-POLICY	33,33%	55,31%	24
Angličtina	CARDINAL-PERSON	60,68%	98,36%	117
	CARDINAL-ORGANIZATION	68,00%	100,00%	25
	ORDINAL-DATE	36,36%	75,75%	11
	ORGANIZATION-DATE	66,67%	100,00%	6
	DATE-ORDINAL	50,00%	50,00%	4

Tabulka 6.13: Tabulka s vypočtenými přesnostmi pro *5* domén s největším počtem nalezených parafrází každého jazyka.

přesnosti uvnitř shluků u obou systémů srovnatelné. Přesnosti vztahů mezi shluky jsou v případě našeho systému nižší což je ovlivněno zejména použitými značkovači.

6.2.4 Nalezení slabých míst systému

Jelikož náš systém vychází z práce [25], lze předpokládat, že bude možno u obou systémů nalézt některá shodná slabá místa. Podle Sekineho jsou hlavní nedostatky systému jednak v tom, že v některých případech dochází k nalézání chybných charakteristických slov z kontextů dvojic pojmenovaných entit a také, že systém vytvoří dvojice pojmenovaných entit i v případech, kdy spolu entity nijak nesouvisí.

V prvním případě, kdy systém nalézá chybná charakteristická slova, dochází buď k tomu, že charakteristické slovo se skládá z více slov, ale systém umí nalézt pro každý kontext pouze jednoslovné charakteristické slovo. Jako příklad Sekine uvádí pojmy *prime minister, vice chairman* nebo *pay for*. Díky tomuto nedostatku jsou pak tyto fráze zařazeny do kategorií *minister, chairman* a *pay*, což následně zapříčiní nižší kvalitu parafrází uvnitř shluků. V našem systému se tomuto jevu částečně vyhýbáme tím, že používáme strukturované parafráze. Sekine navrhuje tuto situaci řešit např. tím, že slova, která se často vyskytují vedle sebe přiřadí k sobě a bude s nimi pracovat tak jako by se jednalo o jedno slovo. V našem systému se s tímto problémem setkáváme např. u slovních spojeních *předseda vlády, předseda sněmovny atd.*.

Dalším problémem je, že systém často nachází charakteristická slova, která nejsou pro danou doménu nijak významná. Sekine tento problém řeší tak, že slova, která mají nízké *tf-idf* skóre, nepovažuje za charakteristická. Tato metoda ovšem neodfiltruje slova, která se nacházejí v doméně často a i přesto nejsou pro danou doménu významná. V našem systému je tato situace řešena pomocí seznamu zakázaných slov, který byl manuálně vytvořen pro oba jazyky. Na tento seznam byla umístěna slova, která by neměla být považována za charakteristická slova v daném jazyce. Toto řešení se zdá být lepší než u Sekineho systému, nicméně seznam zakázaných slov je společný pro všechny domény a proto se může stát, že na něj budou umístěna slova, která pro nějakou doménu významná nejsou, ale pro jinou ano. To vede následně ke snížení počtu nalezených parafrází, jelikož nejsou nacházeny parafráze mezi kontexty, které jsou charakterizovány slovy ze seznamu zakázaných slov jako např. *alias∼čili, kromě∼mimo atd.*

Posledním problémem, který Sekine zmiňuje, je, že systém dává do souvislosti pojmenované entity, které spolu v textu nemají přímou souvislost a nehodí se pro hledání parafrází. Jako příklad pak uvádí větu *Mr. Smith estimates Lotus will make profit this quarter...*, kde systém vyextrahuje pouze *Smith estimates Lotus*, což pak může v některých případech vést k nalezení nekorektních parafrází. V našem systému to pak lze demonstrovat na příkladu *Our local elementary school is considered one of the best around.*, kde systém vyextrahuje *school is considered one*. Pro vyřešení těchto problému je potřeba pro každou větu vytvořit syntaktický strom.

Kromě těchto problémů byly v našem systému objeveny ještě některé další, které se projevují obzvláště při hledání parafrází v češtině. Zejména v doménách, které se týkají politiky, dochází často k jevu, kdy některý politik zastával v různých časových obdobích různé funkce. Jeho jméno je pak často dáváno do souvislosti s těmito funkcemi, což pak vede k tomu, že jsou tyto funkce považovány za parafráze. Příkladem takových chybně nalezených parafrází mohou být např. dvojice *poslanec∼premiér*, *předseda∼ministr*, *ministr∼premiér* či *premiér∼prezident*.

Posledním a zároveň největším nedostatkem, týkajícím se češtiny, je velmi nízká kvalita značkovače pojmenovaných entit. Tento značkovač má problémy při rozpoznávání značného množství víceslovných pojmenovaných entit, a proto nerozpozná např. organizace jako *České dráhy, Česká pošta, Česká spořitelna atd.* nebo některá jména jako např. *Petr Kysela, Jan Hanovec, Martin Bechyňský atd.*. To vede v důsledku k tomu, že z věty *Tiskový mluvčí České pošty Petr Kysela včera oznámil...* systém vyextrahuje pouze úsek *České pošty Petr*, obdobně pak vyextrahuje z podobných vět úseky *České spořitelny Petr* či *Českých drah Petr*, což vede v důsledku k tomu, že slova *dráha*, *pošta* a *spořitelna* jsou pokládány za parafráze. Této situaci nelze zabránit vhodným nastavením parametrů systému. Jediným řešením je vylepšit práci značkovače pojmenovaných entit.

6.2.5 Technické parametry systému

Celý systém je rozdělen do dvou hlavních částí. První část *Paraphrase Discover* slouží k získávání parafrází a druhá *Paraphrase Evaluator* slouží k vyhodnocování přesnosti nalezených parafrází. Obě dvě tyto části jsou napsány v jazyce Java. *Paraphrase Discover* nezískává parafráze přímo z textu, ale již z extrahovaných dvojic pojmenovaných entit. Tyto dvojice musí být uloženy ve formátu *COUPLES* a dají se získat z dat pomocí připravených skriptů. Veškeré parametry systému lze nastavit pomocí konfiguračního souboru *paraphrase.ini*. Popis všech parametrů spolu s ovládáním programu je uveden v dokumentaci k systému, která se nachází na přiloženém elektronickém nosiči. Systém ke svému chodu vyžaduje mít nainstalovány tyto součásti:

- Java[4]; Pro vlastní chod systému.
- Ant[5]; Pro kompilaci a spuštění systému.
- Python[6]; Pro předzpracování dat a extrakci dvojic pojmenovaných entit.
- PDT 2.0[7]; Pro označkování českých textů - volitelné.

[4] http://www.java.com/getjava/
[5] http://ant.apache.org/
[6] http://www.python.org/
[7] http://ufal.mff.cuni.cz/pdt2.0/

V systému lze rovněž pro získávání parafrází použít vlastních dat, která jsou pomocí skriptů označkována a jsou z nich extrahovány dvojice pojmenovaných entit ve formátu *COUPLES*. To lze provést pouze v případě češtiny, kde je možné provést označkování pomocí nástrojů z PDT 2.0. Tato možnost byla do systému zabudována kvůli plánovanému zpracování dat ze serverů *novinky.cz* a *zpravy.idnes.cz*. Systém k chodu nutně nevyžaduje mít nainstalován žádný značkovač (PDT 2.0 ani SuperSense Tagger), protože využívá již označkovaná data. Označkovaná data pro češtinu lze nalézt na školním serveru *merlin* v adresáři */mnt/minerva1/nlp/corpora/monolingual/czech* (soubor *tagged.vert*) a pro angličtinu v adresáři */mnt/minerva1/nlp/corpora/monolingual/english/wikipedia/SW1* (soubory **.txt*). Jelikož tyto soubory dosahují velikosti několika desítek GB, nebudou přikládány na elektronický nosič.

Časová i paměťová náročnost systému je závislá na použitých datech a nastavení mnoha parametrů. Extrahování dvojic pojmenovaných entit trvá na školním serveru *pcnlp1* cca 8–10 hodin při paměťové náročnosti 2–4 GB RAM. Vlastní získávání parafrází ze souborů *COUPLES* pak trvá již jen několik desítek vteřin na tom samém serveru. Paměťová náročnost ovšem dosahuje stále 2–4 GB RAM.

Kapitola 7

Závěr

Úkolem této práce bylo prostudovat přístupy pro automatické vyhledávání parafrází. Při plnění tohoto úkolu bylo přečteno značné množství článků, které byly v minulých letech publikovány v rámci seminářů týkajících se parafrází. Dalším úkolem bylo seznámit se s dosavadními výsledky dosaženými v této oblasti. Jak bylo zjištěno, situace v současné době není uspokojivá a i ty nejlepší systémy pro získávání parafrází nesplňují očekávání, která jsou do nich vkládána. Většina systémů je dnes navíc úzce doménově zaměřena, proto je nelze použít pro širší oblasti. V dalším kroku byly prostudovány některé nástroje, které se dají při tvorbě systémů pro vyhledávání parafrází použít. Jako vhodné nástroje se jeví značkovače pojmenovaných entit. Pro angličtinu těchto nástrojů existuje více a jeden z nich byl v této práci popsán. Co se týče češtiny, nebyl doposud žádný takový nástroj vytvořen. Jsou zde proto uvedeny postupy, které jeho absenci alespoň částečně kompenzují. Dalším úkolem bylo shromáždit data pro testování jednotlivých částí systému. Pro češtinu byl vybrán velký korpus českého jazyka, který byl poskytnut FI MU, pro angličtinu pak korpus projektu Semantically Annotated Snapshot of the English Wikipedia. Oba korpusy byly již označkovány, proto bylo jejich zpracování snadné. V neposlední řadě bylo také úkolem této práce implementovat systém pro vyhledávání parafrází. Jako vzor byl použit systém Satoshi Sekineho, který je popsán v [25]. Tento systém je zaměřen na vyhledávání parafrází mezi kontexty dvou pojmenovaných entit. Posledním úkolem této práce bylo vyhodnotit navržený systém a srovnat jej s podobnými systémy. Náš systém byl vyhodnocen pomocí lidských hodnotitelů a porovnáním se systémem Satoshi Sekineho bylo zjištěno, že oba systémy dosahují srovnatelných výsledků. Náš systém vykazoval horší výsledky pouze při vyhledávání parafrází v českém jazyce, což bylo zapříčiněno zejména použitím méně kvalitního značkovače pojmenovaných entit.

Úkolem této práce nebylo udělat průlom v oblasti vyhledávání parafrází, ale využít dosavadních poznatků a principů k tvorbě systému pro vyhledávání parafrází. Za přínos můžeme považovat vylepšení systému popsaného v [25], zejména pak v zavedení jisté strukturovanosti nalezených parafrází. Přínosem rovněž je použití této metody v češtině. I přesto, že v češtině neexistují vhodné nástroje potřebné k tvorbě takového systému, bylo ukázáno, že lze tyto nástroje do značné míry nahradit.

Možností dalšího vývoje této práce je mnoho. Jako hlavní směry, kterými by se však měl vývoj ubírat, lze zmínit nalezení a použití kvalitnějších dat, ve kterých jsou parafráze vyhledávány. Dalším směrem by měl být také vývoj značkovače pojmenovaných entit pro češtinu. Ačkoliv byl tento značkovač relativně úspěšně nahrazen nástroji z PDT 2.0, s kvalitou značkovačů pro angličtinu se nemůže ani z daleka srovnávat. V neposlední řadě by také měly být vyřešeny další problémy, které byly zmíněny v kapitole 6.2.4, jako např. lepší

identifikace víceslovných charakteristických slov nebo odfiltrování dvojic pojmenovaných entit, které spolu nesouvisí.

Literatura

[1] Association for Computational Linguistics: The 1st International Workshop on Paraphrasing (IWP2001). 2001.
URL http://nlp.nagaokaut.ac.jp/pub/NLPRS2001WS-CFP.html

[2] Association for Computational Linguistics: The 2nd International Workshop on Paraphrasing (IWP2003). 2003.
URL http://nlp.nagaokaut.ac.jp/IWP2003/CFP.html

[3] Association for Computational Linguistics: The 3rd International Workshop on Paraphrasing (IWP2005). 2005.
URL http://nlp.nagaokaut.ac.jp/IWP2005/CFP.html

[4] Bannard, C.; Callison-Burch, C.: Paraphrasing with Bilingual Parallel Corpora. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, Ann Arbor, Michigan, USA: Association for Computational Linguistics, June 2005, s. 597–604.
URL http://www.aclweb.org/anthology-new/P/P05/

[5] Barzilay, R.; Lee, L.: Learning to Paraphrase: An Unsupervised Approach Using Multiple-Sequence Alignment. In *Proceedings of 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT/NAACL 2003)*, Edmonton, Canada: Association for Computational Linguistics, May–June 2003, s. 16–23.
URL http://www.aclweb.org/anthology-new/N/N03/

[6] Barzilay, R.; Mckeown, K. R.: Extracting paraphrases from a parallel corpus. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL 2001)*, Toulouse, France: Association for Computational Linguistics, July 2001, s. 50–57.
URL http://www.aclweb.org/anthology-new/P/P01/

[7] Bhagat, R.; Pantel, P.; Hovy, E.: LEDIR: An Unsupervised Algorithm for Learning Directionality of Inference Rules. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, Prague, Czech Republic: Association for Computational Linguistics, June 2007, s. 161–170.
URL http://www.aclweb.org/anthology-new/D/D07/

[8] Boonthum, C.: iSTART: Paraphrase Recognition. In *Proceedings of the ACL Student Research Workshop (ACL 2004)*, editace L. van der Beek; D. Genzel; D. Midgley,

Barcelona, Spain: Association for Computational Linguistics, July 2004, s. 31–36.
URL http://www.aclweb.org/anthology-new/P/P04/

[9] Brockett, C.; Dolan, W. B.: Support Vector Machines for Paraphrase Identification and Corpus Construction. In *Proceedings of the 3rd International Workshop on Paraphrasing (IWP 2005)*, Jeju Island, South Korea: Association for Computational Linguistics, October 2005, s. 1–8.
URL http://www.aclweb.org/anthology-new/I/I05/

[10] Dagan, I.; Glickman, O.: Probabilistic Textual Entailment: Generic Applied Modeling Of Language Variability. In *Proceedings of PASCAL Workshop on Learning Methods for Text Understanding and Mining*, Grenoble, France, January 2004.
URL http://pascallin.ecs.soton.ac.uk/Workshops/LMTUM04/

[11] Dagan, I.; Glickman, O.; Magnini, B.: The PASCAL Recognising Textual Entailment Challenge. In *Proceedings of the PASCAL Challenges Workshop on Recognising Textual Entailment*, April 2005, s. 1–8.
URL http://pascallin.ecs.soton.ac.uk/Challenges/RTE/

[12] Dolan, W. B.; Brockett, C.: Automatically Constructing a Corpus of Sentential Paraphrases. In *Proceedings of the 3rd International Workshop on Paraphrasing (IWP 2005)*, Jeju Island, South Korea: Association for Computational Linguistics, October 2005, s. 9–16.
URL http://www.aclweb.org/anthology-new/I/I05/

[13] Fujita, A.; Inui, K.: A Class-oriented Approach to Building a Paraphrase Corpus. In *Proceedings of the 3rd International Workshop on Paraphrasing (IWP 2005)*, Jeju Island, South Korea: Association for Computational Linguistics, October 2005, s. 25–32.
URL http://www.aclweb.org/anthology-new/I/I05/

[14] Glickman, O.; Dagan, I.: Identifying Lexical Paraphrases From A Single Corpus - A Case Study For Verbs. In *Proceedings of Recent Advances in Natural Language Processing 2003 (RANLP 2003)*, Borovets, Bulgaria, September 2003.
URL http://lml.bas.bg/ranlp2003/

[15] Hajič, J.; Hajičová, E.; Hlaváčová, J.; aj.: Průvodce PDT 2.0. 2006.
URL http://ufal.mff.cuni.cz/pdt2.0/doc/pdt-guide/cz/html/index.html

[16] Hana, J.; Zeman, D.: Manual for Morphological Annotation. 2005.
URL
http://ufal.mff.cuni.cz/pdt2.0/doc/manuals/en/m-layer/pdf/m-man-en.pdf

[17] Hasegawa, T.; Sekine, S.; Grishman, R.: Discovering Relations among Named Entities from Large Corpora. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics (ACL 2004)*, Barcelona, Spain: Association for Computational Linguistics, July 2004, s. 415–422.
URL http://www.aclweb.org/anthology-new/P/P04/

[18] Lin, D.; Pantel, P.: DIRT: Discovery of Inference Rules from Text. In *Proceedings of 7th ACM SIGKDD International Conference on Knowledge Discovery and Data*

Mining (KDD 2001), San Francisco, California, USA, August 2001, s. 323–328.
URL http://www.sigkdd.org/kdd2001/

[19] Manning, C. D.; Raghavan, P.; Schütze, H.: *An Introduction to Information Retrieval*. 2009.
URL http://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf

[20] Massimiliano, C.; Yasemin, A.: Broad-Coverage Sense Disambiguation and Information Extraction with a Supersense Sequence Tagger. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, Sydney, Australia: Association for Computational Linguistics, July 2006, s. 594–602.
URL http://acl.ldc.upenn.edu/W/W06/

[21] Pang, B.; Knight, K.; Marcu, D.: Syntax-based Alignment of Multiple Translations: Extracting Paraphrases and Generating New Sentences. In *Proceedings of 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT/NAACL 2003)*, Edmonton, Canada: Association for Computational Linguistics, May–June 2003, s. 102–109.
URL http://www.aclweb.org/anthology-new/N/N03/

[22] Pasca, M.; Dienes, P.: Aligning Needles in a Haystack: Paraphrase Acquisition Across the Web. In *Proceedings of 2nd International Joint Conference on Natural Language Processing (IJCNLP 2005)*, Jeju Island, South Korea: Association for Computational Linguistics, October 2005, s. 119–130.
URL http://www.aclweb.org/anthology-new/I/I05/

[23] Power, R.; Scott, D.: Automatic generation of large-scale paraphrases. In *Proceedings of the 3rd International Workshop on Paraphrasing (IWP 2005)*, Jeju Island, South Korea: Association for Computational Linguistics, October 2005, s. 73–79.
URL http://www.aclweb.org/anthology-new/I/I05/

[24] Qiu, L.; Kan, M.-Y.; Chua, T.-S.: Paraphrase Recognition via Dissimilarity Significance Classification. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP 2006)*, Sydney, Australia, July 2006, s. 18–26.
URL http://www.cs.jhu.edu/~yarowsky/SIGDAT/emnlp06.html

[25] Sekine, S.: Automatic Paraphrase Discovery based on Context and Keywords between NE Pairs. In *Proceedings of the 3rd International Workshop on Paraphrasing (IWP 2005)*, Jeju Island, South Korea: Association for Computational Linguistics, October 2005, s. 80–87.
URL http://www.aclweb.org/anthology-new/I/I05/

[26] Shinyama, Y.; Sekine, S.: Paraphrase acquisition for information extraction. In *Proceedings of the 2nd International Workshop on Paraphrasing (IWP 2003)*, Sapporo, Japan: Association for Computational Linguistics, July 2003, s. 65–71.
URL http://www.aclweb.org/anthology-new/W/W03/

[27] Szpektor, I.; Tanev, H.; Dagan, I.; aj.: Scaling Web-Based Acquisition of Entailment Relations. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP 2004)*, Barcelona, Spain, July 2004, s. 41–48.
URL http://www.cs.ualberta.ca/~lindek/emnlp04/

[28] Wan, S.; Dras, M.; Dale, R.; aj.: Towards Statistical Paraphrase Generation: Preliminary Evaluations of Grammaticality. In *Proceedings of the 3th International Workshop on Paraphrasing (IWP 2005)*, Jeju Island, South Korea: Association for Computational Linguistics, October 2005, s. 88–95.
URL http://www.aclweb.org/anthology-new/I/I05/

[29] Zaragoza, H.; Atserias, J.; Ciaramita, M.; aj.: Semantically Annotated Snapshot of the English Wikipedia v.1 (SW1). 2007.
URL http://www.yr-bcn.es/semanticWikipedia

[30] Zhang, Y.; Patrick, J.: Paraphrase Identification by Text Canonicalization. In *Proceedings of the Australasian Language Technology Workshop 2005 (ALTW 2005)*, Sydney, Australia, December 2005, s. 160–166.
URL http://www.alta.asn.au/events/altw2005/

[31] Ševčíková, M.; Žabokrtský, Z.; Krůza, O.: Zpracování pojmenovaných entit v českých textech. 2007.
URL http://ufal.mff.cuni.cz/~zabokrtsky/reports/techrep-ne-2007.pdf

Dodatek A

Přílohy

Příloha 1: Datový nosič CD se zdrojovými kódy, programovou a uživatelskou dokumentací, použitými daty, souborem README, konfiguračním souborem, plakátem a elektronickou kopií této technické zprávy.