Disambiguace mnohoznačných slov pomocí vizuální informace

Michal Vavrečka, Lenka Lhotská

Biodat, FEL ČVUT
Karlovo náměstí 13, Praha 1
vavrecka@fel.cvut.cz,

Abstrakt

V příspěvku se zabýváme analýzou současných přístupů k identifikaci správného významu mnohoznačných slov (disambiguaci) a návrhem architektury umožňující zlepšit rozlišování významu slov v textových dokumentech nebo obrazových datech. Nejprve se věnujeme základnímu členění klasických metod využívajících k disambiguaci pouze lexikální kontext. Následně představíme systémy, které používají k disambiguaci také obrazovou informaci. V závěru integrujeme oba přístupy do jednotného modelu, který využívá lingvistický, vizuální a prostorový kontext pro identifikaci správného významu slova. Navržený systém vychází z multimodální architektury používané v oblasti reprezentace prostorových vztahů (Vavrečka, 2008, 2010, v tisku).

Klíčová slova: disambiguace, polysémy, homonyma, multimodální reprezentace, samoorganizace, WSD

1 Úvod

V oblasti přirozeného jazyka může slovo či fráze zastupovat více významů. Tato schopnost lexikální jednotek mít více významů se nazývá mnohoznačnost nebo polysémie. Příkladem může být slovo *oko*, které znamená lidský orgán, smyčku na chytání zvěře popřípadě díru na punčoše. U člověka je schopnost odhadnout správný význam slova založena na automatických nevědomých mechanismech. Při disambiguaci využíváme co*mmon sense* znalostí, jež jsou pro oblast automatické analýzy založené na algoritmech umělé inteligence velmi obtížně formalizovatelné a proto je tato oblast velkou výzvou pro budoucí výzkumy.

Jelikož je většina současných výzkumů zaměřena na analýzu dokumentů v angličtině, zaměříme se v příspěvku právě na tento jazyk. Na začátek uvádíme stručnou statistiku četnosti jednotlivých slov. Oxfordský anglický slovník obsahuje přibližně 301.100 hlavních hesel (McCum et al., 1992). Celkem katalogizuje asi 500.000 slov, přičemž dalších 500.000 vědeckých pojmů není zařazeno. Největší třídu víceznačných slov tvoří polysémy. Jedná se o slova nebo fráze s různými, ale příbuznými významy (např. *kolo* jako dopravní prostředek nebo jako část dopravního prostředku). Statistika polysémie v anglickém jazyce je dostupná pomocí WordNetu. Jedná se o velkou digitální databázi podobnou sémantické síti reprezentující vztahy mezi slovy, včetně synonymie, antonymie, generalizace, lokalizace a specifikace. Jednotlivé pojmy jsou hierarchicky reprezentované (hyperonyma a hyponyma) a také sdružovány podle sémantických vlastností. Existují zde *synsety*, které zastupují jeden z významů polysémních slov. Celá databáze vytváří speciální znalostní bázi zachycující vztahy mezi jednotlivými slovy a je vhodným nástrojem pro automatickou textovou analýzu a disambiguaci. WordNet je nejčastěji používaná databáze v oblasti zpracování přirozeného jazyka. WordNet 3.0. obsahuje 155,287 slov, přičemž 17 procent jsou polysémy. Nejčastěji se jedná o podstatná jména (15.595), která tvoří 60% všech polysémních slov. Průměrná polysémie (počet významů na slovo) ve WordNetu je 2.89, vyloučíme-li slova mající pouze jeden význam (Fehlbaum, 1998).

Přechod od polysémie k homonymii je plynulý. Homonyma jsou podmnožinou polysémů. Jedná se o skupinu slov majících podobný pravopis (homografy), popřípadě stejnou výslovnost (homofony), ale různé a nesouvisející významy, tzn. nemají žádný společné sémantické rysy, ze kterých by bylo možné odvodit společný základ, např. slovo *prát* ve významu bít se nebo jako proces čistění prádla. V užším významu se výraz homonymum vztahuje pouze na slova se stejným pravopisem (homografy). V češtině jsou homonyma poměrně řídká a jejich užívání většinou nezpůsobuje nedorozumění. Naopak v anglickém jazyce se homonyma vyskytují velmi často. Slovník homonym obsahuje 8800 záznamů (Burke, 2009) a 5,879 z nich jsou homografy. Zajímavá je statistika týkající je průměrného počtu významů na homonymum. Slovník obsahuje pouze 3 pojmy (pokud vyloučíme užití množného čísla), které mají více než 10 různých významů (jedná se o anglická slova *point*, *set* a *snap*). Ve slovníku je dále 267 slov s více než 5 významy a 1925 slov s více než 3 významy. Průměrná homonymie (počet významů na slovo) je 2,2.

2 Lexikální disambiguace

V obecné rovině sestává proces lexikální disambiguace ze dvou základních kroků. Nejprve se stanoví všechny významy daného slova v textu a následně se určí nejvhodnější smysl na základě jeho kontextu. V oblasti umělé inteligence se pro automatickou identifikace správného významu slova vžil termín *word sense disambiguation* (WSD). Metody WSD mají více než šedesátiletou historii. První pokusy o automatickou disambiguaci byly ovlivněny Chomského syntaktickou teorií (1957), ale ukázalo se, že tento přístup není příliš efektivní a nedosahuje v oblasti disambiguace uspokojivých výsledků. Během dalšího vývoje byly často používány sémantické sítě (Collins, Quillian, 1963) popřípadě rámce (Hayes, 1976), tedy znalostní báze sloužící k reprezentací vztahů mezi jednotlivými slovy. Metody založené na znalostních bázích využívají různých druhů slovníků a tezaurů. Mezi nejčastější znalostní systémy patří WordNet (Fellbaum, 1998), Open Mind Word Expert (Chklovski a Mihalcea, 2002), Wikipedia (Mihalcea, 2007) nebo paralelní korpusy (Ng et al., 2003). Takto založené systémy porovnávají neznáme slovo se znalostní bázi, přičemž berou v potaz ostatní slova v testovaném dokumentu (kontextové okno), jež porovnávají s významy definovanými ve znalostních bázích. Yarowsky v letech 1993-1994 prokázal, že velikost kontextového okna může mít vliv na kvalitu disambiguace. Dochází k závěru že pro lexikální (syntaktickou) analýzu dostačuje 3-4 slov v kontextovém okně a pro sémantickou (významovou) disambiguaci je potřeba v průměru 50 slov. Neexistuje ovšem lineární vztah mezi velikostí kontextového okna a přesnosti disambiguace. Gale et al. (1993) zjistil že jestliže zvětšíme kontextové okno z 12 na 100 slov, stoupá přesnost disambiguace pouze o 4 procenta. V současnosti bývá pro porovnávání použita široká škála algoritmů (Leskův algoritmus (Lesk, 1986), pravděpodobnostní statistiky, sémantické hierarchické struktury, sémantické podobnosti vypočtené ze sémantické sítě, omezující pravidla, heuristiky apod.).

Metody založené na znalostních bázích dosahují v průměru 60 % spěšnost při disambiguaci polysémů (testováno na korpusu SensEval 3), ale malá přesnost může být způsobena nedostatečné popisy některých významů slov ve znalostní bázi.

Úspěšnost WSD systémů také závisí na typu učení. Systémy založené na učení s učitelem využívající ručně anotovaných příkladů významu každého slova, přičemž přesnost pro homonyma je kolem 80% (Navigli et al. 2007). Nejlepších výsledků dosahují metody založené na strojovém učení, konkrétně za použití *support vector machines* (SVM). Nevýhodou takových systémů je nutnost manuálně anotovat všechny významy slov, což je časově náročné, a proto systémy selhávají, pokud jsou použity na rozsáhlé dokumenty (velké objemy textových dat). Metody založené na učení bez učitele (také známý jako *word sense discrimination*) dokáží pracovat s neanotovanými korpusy. Významy jsou z textu odvozovány na základě míry podobnosti (Lin, 1997). Vycházejí z předpokladu že jednotlivé významy se vyskytují v podobných kontextech, a proto stačí systém natrénovat na velkém množství neanotovaných dokumentů. Jednotlivé významy budou na základě odlišných kontextů vytvářet oddělené clustery. Přesnost těchto systému v oblasti disambiguace homonym je kolem 60% (Navigli et al., 2007). Úspěšnost rozlišení polysémů bývá nižší, nejjednodušší algoritmy dosahují efektivity kolem 50%. Polysémní slova jsou na rozdíl od homonym v některých případech obtížně rozlišitelná i pro člověka. Například při anotaci korpusu Senseval-2 se posuzovatelé shodli pouze v 85% případů.

Využití systému založených na učení bez učitele je velkou výzvou pro příští výzkum, neboť internet obsahuje obrovské množství neanotovaných textových dat vhodným k učení. Tento způsob umožňuje překonat problémy související s ruční anotací.

3 Vizuální disambiguace

Vědci na konferenci SemEval dospěli k závěru, že tradiční metody WSD založené na lexikální analýze dosáhly stádia, kdy nelze příliš zlepšit jejich efektivitu (Aggire, Edmonds, 2007). Nabízí se tedy možnost obohatit tyto architektury o další zdroj informací, který by mohl zlepšit výsledky disambiguace. Inspiraci můžeme hledat například v oblasti reprezentace znalostí založené na integraci informací z různých modalit (např. vizuální), ve kterých se kombinují sekvenčně zpracovávané symbolové informace (jazyk) s informacemi zpracovávanými paralelně (obraz). Předběžné pokusů v této oblasti (Barnard, Johnson, 2005; Saenko, Darrell, 2008) vedly k 10-15% zvýšení přesnosti v porovnání s tradičními metodami WSD.

Z teoretického hlediska je problémem čistě lexikálních systémů neschopnost zpracovávat informace o typických vlastnostech popsaných objektů, jako je tvar, barva, velikost, prostorové a funkční vlastnosti, které jsou nezbytné pro přesnou disambiguaci. Tato nevýhoda je v oblasti reprezentace znalostí známá jako *symbol grounding problem* (Harnad, 1990). Proto se nabízí možnost obohatit klasický WSD systém o informace z vizuálních domény. Polysémní slova jsou v takovém případě disambiguovány nejen pomocí jiných slov v rámci lexikálního kontextu, ale také vizuálními a prostorovými vlastnostmi. Příkladem může být anglické homonymum klíč (*key*), které má 5 odlišných významů (Burke, 2009). Tři z pěti významů jsou definovány jako „nástroj k odemknutí“, „nástroj pro uchopení matek“ a „klávesa na klavíru/varhanách“. Cílem je rozpoznat

správný význam slova klíč (*key*) v kontextovém okně „Židle stojí před nástrojem se spoustou kláves (*keys*). Některé z nich jsou černé a většina z nich bílé.“. V tomto případě selžou metody založené na analýze lexikálního kontextu, jelikož systém na základě dostupných informací rozpozná klíč jako *nástroj odemykání* nebo *nástroj pro uchopení matek*. Pokud ovšem použijeme k disambiguaci také sybsymbolové znalosti, může systém procházet reprezentační bází obsahující abstrahovanou informaci o obrázcích na kterých se objevuje jednotlivé významy slova klíč (*key*). V případě významu „klávesa na klavíru/varhanách“ se jedná o obrázky klavíru či varhan v různých situacích. Ve většině případů bývá klavír na obrázku společně se židlí a tuto informaci reprezentujeme ve specializovaném subsystému (tvarové a prostorové vlastnosti objektů). Pokud zároveň použijeme subsystém, který uchovává prototypické informace o barevných vlastnostech popisovaných objektů, dokáže systém správně přiřadit správný význam slova klíč (*key*) jako páky (klávesy) klavíru. Použitím subsystémů pro reprezentaci typických prostorových vztahů a barev zlepšíme nízkou přesnost samostatně aplikovaných algoritmů pro rozpoznávání obrazu nebo lexikální analýzu. Ukázkou může být níže popsaný návrh systému zpracovávajícího lingvistický, vizuální, prostorový a barevný kontext polysémů a dokáže jednotlivé informace propojit dohromady, čímž zvyšuje přesnost disambiguace.

4 Návrh modelu

Námi navrhovaná metoda je založená na učení bez učitele, kombinující ligvistické informace s vizuální a prostorovou kontextovou informací. Základní myšlenka spočívá v propojení klasického systému založeného na lexikální analýze kontextu s algoritmy pro rozpoznávání vizuálních a prostorových vlastností objektu. Systém může být použit jako součást automatického anotátoru neznámých dokumentů, v oblasti strojového překladu, v internetových vyhledávačích (jak textových i grafických), jako součást expertního systému nebo nový typ znalostní báze.

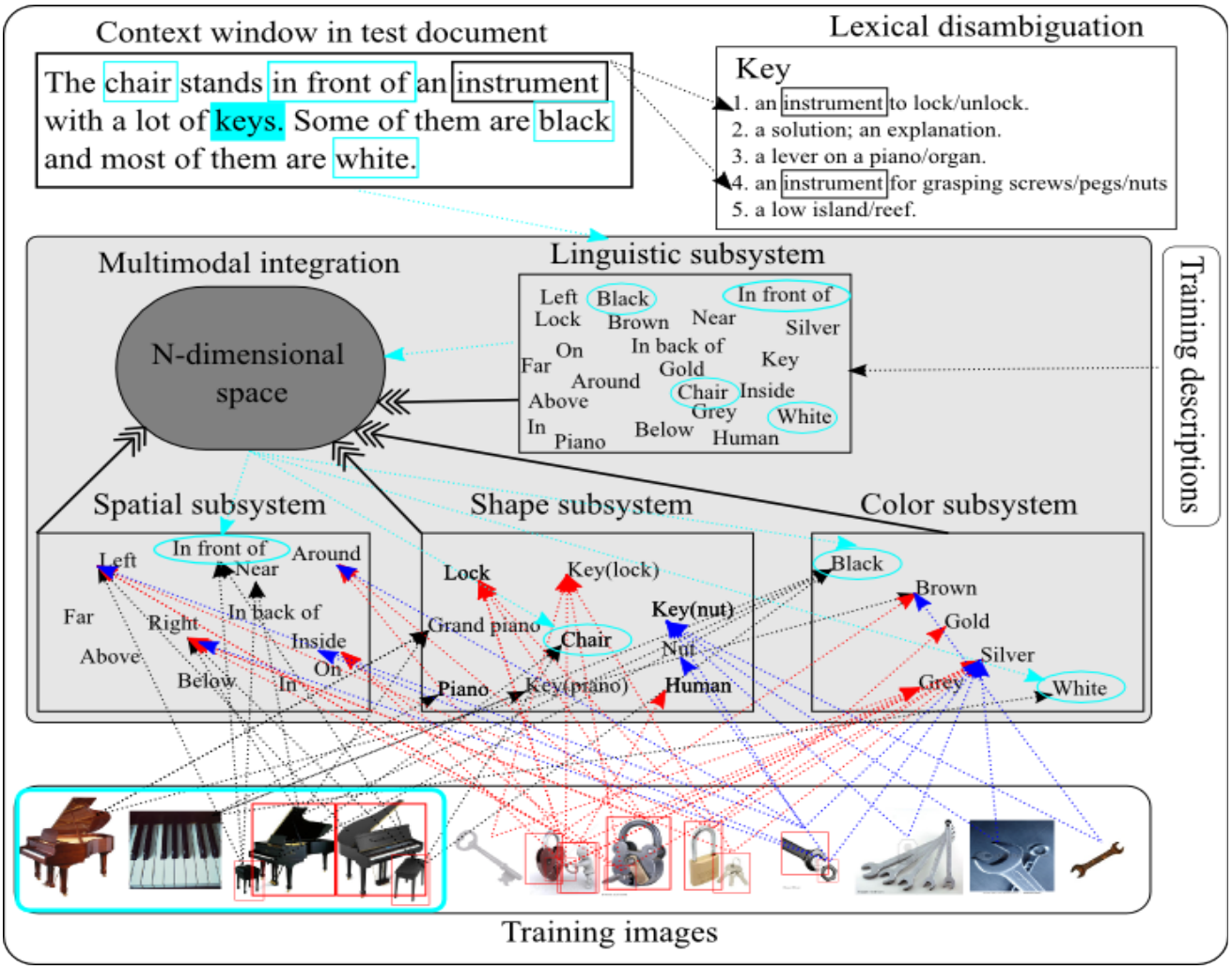
Navrhovaný model kombinuje klasické metody pro WSD s metodami pro subsymbolovou reprezentaci prostorových a vizuálních informací. Doplnění o prostorový subsystém je založeno na předpokladu, že společný výskyt více objektů ve scéně reprezentuje jejich prostorové a funkční vztahy. Jestliže se objekty vyskytují blízko sebe (např. židle a klavír), předpokládáme, že spolu nějak souvisí. Jestliže tuto znalost reprezentujeme v subsystému prostorových vztahů, může nám to pomoci při nalezení správného významu slova klávesa (*key*), jestliže se v neanotovaném dokumentu objeví v blízkosti slova židle.

Samotný způsob fungování navrhovaného systému se dá popsat v následujících krocích. Nejprve je potřeba natrénovat systém na známých datech. V této fázi se systém učí nalézt mapování mezi vizuálními vlastnostmi (tvar, barva, prostorový vztah) a textovými vstupy. Pro tyto účely je použit slovník či databáze, která obsahuje seznam homonym či polysémů. Ideálním kandidátem je Wordnet (Fellbaum, 1998), díky dobré strukturovanosti polysémních slov. Pro získaní vizuálních dat, které odpovídají jednotlivým významům polysémů je vhodná databáze Image-Net (Deng et al., 2009). Jedná se veřejně dostupnou obrazovou banku, která je setříděna podobně jako jednotlivé výrazy ve WordNetu a navíc je vybavena vlastním API pro efektivní manipulaci s daty. Obsahuje 1,2 milionu obrázků (1000 obrázků na *synset*), přičemž všechny obrázky obsahují SIFT (scale invariant feature transformations) popisující jejich lokální invariantní vlastnosti. Polovina obrázků v databází (650.000) je navíc doplněna o anotované výřezy obsahující jednotlivé objekty ve scéně. Abychom mohli reprezentovat informace z těchto vstupů, je nutné použít adekvátní architekturu. Vhodným kandidátem může být multimodální architektura založená na učení bez učitele (Vavrečka, 2008, 2009, 2010, v tisku), která dokáže integrovat subsymbolovou informaci (vizuální a prostorové vlastnosti) s informací symbolovou (textové vstupy). Tato architektura byla odzkoušena v oblasti reprezentace prostorových vztahů, přičemž je možné využít její potenciál i v oblasti disambiguace. Rozšířením vznikne systém obsahující primární subsystémy pro zpracování informace o tvaru objektu (vstupem jsou výřezy obrázků nebo celé obrázky), barvě (vstupem je kombinace nejvíce zastoupených barev v daném obrázku), prostorových vztazích (vstupem je zjednodušené schéma polohy objektů ve scéně), invariantních vlastnostech ve formě SIFT vektorů a symbolové (textové) informaci (vstupy jsou jednoduché popisy scény). Popis scény (prostorové vztahy, barvu a názvy objektů) je možné automaticky extrahovat z obrázků obsahujících výřezy více objektů nebo poloautomaticky z obrázků obsahujících více objektů ve scéně.

Tyto základní subsystémy je možno nazírat jako specifické "sémantické sítě" reprezentující strukturu a vztahy v rámci lingvistického, prostorového a vizuálního (tvar a barva) kontextu. V dalším kroku je výstup z uvedených subsystémů projikován do multimodální vrstvy, která integruje jednotlivé vstupy do jednotné koherentní mapy. Ta slouží jako "překladač" z jednoho subsystému do ostatních. Jsme proto schopní například pomocí lingvistického vstupu identifikovat odpovídající clustery nejen v symbolovém (lingvistickém) subsystému, ale také v ostatních subsystémech (viz. Obr.1). Multimodální vrstva slouží jako abstraktní vrstva sémantické sítě propojující lokální kontexty do jediného

celku, čímž doplňuje chybějící vztahy mezi objekty v

Nový krok ve zpracování spočívá ve využití vizuální báze



Obr. 1. *Schéma systému. Během tréninku jsou systému prezentovány vizuální a lingvistické vstupy, které integruje ve společné multimodální vrstvě. Během testování jsou předkládány nové data, které aktivují odpovídající clustery v jednotlivých subsystémech.*

jednotlivých subsystémech.

Proces testování navrhovaného systému na nových datech lze provádět následujícím způsobem. V první fázi je činnost podobná klasickým lexikálním WSD systémům. Neznámý dokument je prohlédnut algoritmem, který detekuje polysémy na základě porovnání s lexikální databází (např. Wordnet). Věty jsou segmentovány pomocí interpunkce (např. *?,!,.*) a poté je provedena analýza větných členů (POS) a dalších prvků upřesňujících strukturu dokumentu. Následuje klasický proces WSD založený na prohledávání kontextového okna, identifikaci klíčových slov a jejich porovnávání se znalostní bází (Wordnet, Wikipedia, CYC atd.) vedoucí k disambiguaci polysémních slov.

výše popsaného systému, pro disambiguaci slov, které nedokázal lexikální algoritmus správně rozlišit. V kontextovém okně jsou kromě klíčových slov lokalizovány také slova popisující prostorové vztahy a barvy a jejich vztah k homonymům a kontextových slovům. Informace o prostorových vztahů (pokud je přítomna), barvě (pokud je přítomna), homonymu, kontextových slovech (případně jejich synonymem, hyperonymech a hyponymech), bude sloužit jako vstup pro lingvistickou část systému a multimodální vrstva následně aktivuje příslušné clustery v ostatních subsystémech. Navrhovaná architektura dokáže navíc v případě absence popisu prostorových vztahů tuto informaci rekonstruovat pomocí kontextových slov

(pokud byly přítomné ve stejné scéně během tréninku). Pro správné určení významy slova slouží speciální algoritmus, který porovná vzdálenosti jednotlivých clusterů v subsystémech. Nastavení správných vah pro jednotlivé subsystémy nelze apriorně odhadnout, proto je bude nutné otestovat na reálných datech. Jestliže se prokáže vyšší efektivita systému v porovnání s klasickou lexikální analýzou, může se navrhovaný systém stát součástí automatického anotátoru textových dokumentů, algoritmu pro strojový překlad popřípadě internetového vyhledávače.

Pro hodnocení efektivity WSD systému se nejčastěji používá standardizovaný korpus, např. SemEval. Ten se vyvíjel v letech 1998-2010 od verze Senseval-1 až po současnou variantu SemEval2 a obsahuje speciální úlohy věnované disambiguaci polysémních slov. Testovacích data se skládají ze tří dokumentů (6000 slov z nichž je přibližně 2.000 víceznačných), které jsou ručně anotované pomocí více hodnotitelů. Míra úspěšnosti testovaného algoritmů (systémů) se vyjadřuje pomocí dvou škál. Jedná se o *precision* zastupující poměr počtu správně označených slov k celkovému počtu slov zpracovaných systémem a *recall* reprezentující počet slov označených správně v poměru k počtu slov v testovací množině. Někteří výzkumníci používají pro vyjádření efektivity hodnotu F, která je průměrem mezi *precision* a *recall*. Pokud například systém analyzuje 75 slov z testovací sady obsahující 100 slov a určí správně 50 slov, pak je hodnota *precision* 50/75 = 0,66, hodnota *recall* je 50/100 = 0,50 a hodnota F odpovídá 0,58.

5 Závěr

Navrhovaný model má demonstrovat možnosti reprezentace znalostí rozdílných typů informací a jejich vzájemné mapování. V současné fázi bylo prokázáno, že je systém schopný mapovat jednotlivé slova věty na příslušné subsymbolové reprezentace (Vavrečka, in press). Během následující fáze bychom rádi otestovali navrhovaný model na reálných datech a určili jeho efektivitu v oblasti disambiguace mnohoznačných slov.

Poděkování

Tento projekt byl financován z výzkumného záměru MSM 6840770012 a Národního stipendijního programu SAIA pod názvem „Multimodální reprezentace v oblasti vnímání prostoru“.

Literatura

[1] E. Aggire, P.G. Edmonds, *Word Sense Disambiguation: Algorithms and Applications (Text, Speech and Language Technology)*, Springer, 2007.

[2] K. Barnard, M. Johnson,Word Sense Disambiguation with Pictures, *Artificial Intelligence, Volume 167*, pp. 13-30, 2005.

[3] R. R. Burke, *Dictionary of Homonyms a Homophones.* 7th edition, 2009.

[4] A. M. Collins, M.R. Quillian, Retrieval time from semantic memory. *Journal of verbal learning and verbal behavior* 8 (2): 240–248, 1969.

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, ImageNet: A Large-Scale Hierarchical Image Database. *IEEE Computer Vision and Pattern Recognition (CVPR)*, 2009.

[6] Ch. Fellbaum, *WordNet: An Electronic Lexical Database.* Cambridge, MA: MIT Press, 1998.

[7] W.A. Gale, K.W. Church, D. Yarowsky, A Method for Disambiguating Word Senses in a Large Corpus, *Comput. Humanities*,vol. 26, pp. 415– 439, 1993.

[8] S. Harnad, The Symbol Grounding Problem. *Physica D,* 335-346, 1990.

[9] P.J. Hayes, A Process to Implement Some Word Sense Disambiguation, Working Paper 23, Institut pour les Etudes Semantiques et Cognitives, Universite de Geneve, 1976.

[10] H. T. Ng, B. Wang, Y. S. Chan, Exploiting Parallel Texts for Word Sense Disambiguation: An Empirical Study, In *ACL*, 2003 .

[11] T. Chklovski, R. Mihalcea, Building a sense tagged corpus with open mind word expert. In *Proceedings of the Acl-02 Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*, Morristown, NJ, 116-122, 2002.

[12] N.Chomsky, *Syntactic Structures*, The Hague: Mouton, 1957.

[13] M. Lesk, Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone, ACM Special Interest Group for Design of Communication,In *Proc. of the 5th Ann. Int. Conf. on System Documentation*, 1986, pp. 24– 26.

[14] D. Lin, Using syntactic dependency as local context to resolve word sense ambiguity. *In Proceedings of the 35th Annual Meeting of the Association For Com-putational Linguistics and Eighth Conference of the European Chapter of the Association For Computa-tional Linguistics,* Madrid, Spain, July 07 - 12, 1997.

[15] R. McCrum, William Cran, Robert MacNeil, *The Story of English*. New York: Penguin, 1992.

[16] R. Mihalcea, P. Edmonds, *Proc. of Senseval 3: The Third Int. Workshop on the Evaluation of Systems for the Semantic Analysis of Text,* Barcelona, Spain, 2004.

[17] R. Mihalcea, Using Wikipedia for Automatic Word Sense Disambiguation, In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL 2007)*, Rochester, April , 2007.

[18] R. Navigli, K. Litkowski, O. Hargraves, SemEval-2007 Task 07: Coarse-Grained English All-Words Task. Proc. of Semeval-2007 Workshop (SemEval), In *45th Annual Meeting of the Association for Computational Linguistics (ACL 2007*), Prague, Czech Republic, pp. 30-35, 2007

[19] K. Saenko, T. Darrell, Unsupervised Learning of Visual Sense Models for Polysemous Words . *Proc. NIPS*, December 2008.

[20] M. Vavrečka, Application of the cognitive semantics in the model for the spatial terms representation. *Unpublished PhD thesis, Masaryk University in Brno, Czech Republic,* 2008.

[21] M. Vavrečka, I. Farkaš, Unsupervised model for grounding multimodal representations. *Third EuCogII Members Conference, Mallorca,* 2010

[22] M. Vavrečka, I. Farkaš, Unsupervised Grounding of Spatial Relations. In *Proceedings of European Conference on Cognitive Science*, Sofia, Manuscript submitted for publication.

[23] D. Yarowsky, One Sense per Collocation, HLT'93: *Proc. of the Workshop on Human Language Technology*, Morristown, NJ, USA: Association for Computational Linguistics, 1993, pp. 266–271.

[24] D. Yarowsky, Decision Lists for Lexical Ambiguity Resolution: Application to Accent Restoration in Spanish and French, In *Proc. of the 32nd Ann. Meeting of Association for Computational Linguistics,* Morristown, NJ, USA: Association for Computational Linguistics, 1994, pp. 88–95