PAPEL

Palavras Associadas Porto Editora Linguateca

Extracção de relações a partir de dicionários: Breve história

Hugo Oliveira, Paulo Gomes, Nuno Seco
Linguateca, pólo de Coimbra, DEI - FCTUC, CISUC
Diana Santos
Linguateca, pólo de Oslo, SINTEF ICT

Agosto 2008

Índice

1 Introdução

Depois de no primeiro relatório [GGS07] terem sido apresentados alguns recursos semelhantes àquele que pretendemos construir e também as relações que queremos incluir no PAPEL, neste segundo relatório apresentamos o estado da arte no que diz respeito à extracção de informação semântica a partir de um dicionário electrónico.

2 História

Dividimos a história da extracção de relações a partir de dicionários electrónicos em três períodos: o inicial, na década de 70, um segundo período, a década de 80, em que se assistiu a uma maior teorização do assunto, e a década de 90 até aos nossos dias, em que a análise passou a usar ferramentas computacionais mais potentes.

2.1 O início

A utilização de dicionários electrónicos no processamento de linguagem natural data já da década de 1970, com vários trabalhos de Nicoletta Calzolari, onde se começam a explorar as definições de dicionários e a procurar resolver problemas como a circularidade nelas presente [Cal77]. A resolução destes problemas facilita a organização de um dicionário numa base de dados lexical, onde será possível aceder directamente a toda a informação contida nas definições [Cal82]. Estando esta base de dados bem estruturada, torna-se mais simples a identificação automática de algumas relações sintácticas e semânticas entre as várias entradas do dicionário. Tirando partido do vocabulário restrito e específico presente num dicionário Calzolari propõe a detecção de padrões indicadores dessas relações que poderão ser utilizados na sua extracção [Cal84].

Para a língua inglesa, já dos finais da década de 1970 começaram ser realizados trabalhos sobre os dicionários *Longman Dictionary of Contemporary English* (LDOCE), *Merriam-Webster Pocket Dictionary* (MPD) e *Websters 7th Collegiate Dictionary* (W7), existindo um número considerável de publicações ao longo da década de 1980.

Em 1980, Michiels [MMN80], publica um artigo onde são explorados o ficheiros que fazem parte do LDOCE e apresenta a usa estrutura e algumas propriedades características das suas definições. Tal como outros autores, Michiels

chega à conclusão que o vocabulário presente num dicionário é bastante limitado, o que pode facilitar o seu processamento na busca de relações entre estruturas sintácticas ou semânticas.

Pela mesma altura, Amsler [Ams80] dissertava acerca da estrutura da versão electrónica do MPD.

Na sua tese refere que a esmagadora maioria das definições segue uma estrutura onde está presente um *genus* e uma *differentia*. O *genus* identifica normalmente o conceito superordinado da palavra definida, ou por outras palavras, diz-nos que o conceito definido é "um tipo de" outro, existindo por tanto uma relação de hiponímia para esse o outro. A *differentia* é a parte da definição responsável pela distinção entre a instância do conceito superodinado das restantes instâncias, através de propriedades mais específicas da palavra definida. Extraindo e desambiguando os *genus* é possível construir hierarquias semânticas baseadas na relação de hiperonímia (no caso dos nomes) e troponímia (no caso dos verbos). Estes termos são utilizados na maior parte das publicações da área.

Acreditando que era possível extrair uma enorme quantidade de informação semântica através do dicionário Amsler propôs uma taxonomia constituída por hierarquias de nomes e hierarquias de verbos (*tangled hierarchies*), construídas após a análise das definições do MPD, baseando-se no núcleo (desambiguado manualmente) de cada definição [Ams81]. As hierarquias foram organizadas de forma a que as palavras mais específicas se encontrassem nos níveis mais baixos e as mais genéricas (como "causa", "coisa", "classe", "ser"...) no topo. São também referidos alguns problemas que surgiram na realização do trabalho, como o já referido por Calzolari problema da circularidade nas definições. Outro problema referido está relacionado com nomes que se encontram definidos através do argumento de verbos ou através de de um todo do qual fazem parte (folha - parte de uma planta), em vez de um termo superordinado (hiperónimo).

2.2 Década de 1980

Em 1985, Chodorow propunha duas heurísticas para identificar o conceito superordinado nas definições em dicionários. Para isso tirou partido do estilo algo previsível que as definições apresentam, não necessitando de efectuar o *parsing* completo de cada uma. Tendo em conta que o conceito definido é normalmente um hipónimo do conceito superordinado, Chodorow utilizou as heurísticas definidas para construir árvores taxonómicas de uma forma recursiva e semi-automática. É contudo necessária a intervenção humana para decidir se uma nova palavra está a ser correctamente inserida na taxonomia

garantindo assim a obtenção de uma árvore desambiguada.

Em 1986, Markowitz propôs um conjunto de padrões de texto que ocorrem no início das definições de um dicionário (W7) e que: 1) indicam relações entre nomes (nomeadamente relação de superordinação e membro-de); 2) que o nome definido representa um ser humano; 3) identificam os verbos ou adjectivos como activos (*active*) ou de estado (*stative*) [MAE86].

Em 1987, Alshawi mostrou efectuou uma análise das definições do LDOCE onde identificou vários padrões sintácticos que possibilitam a construção de estruturas semânticas baseadas nos significados definidos [Als87]. As estruturas semânticas são derivadas a partir da identificação dos termos subordinados ou de modificadores, preposições e outras palavras que possam indicar relações que estejam presentes na definição. As estruturas são constituídas por um conjunto de relações semânticas e em alguns casos propriedades características das mesmas. Hiponímia (*class*), objectivo (*purpose*), forma (*manner*) ou parte (*has-part*) são apenas algumas das relações presentes nas estruturas.

2.3 Década de 1990 até à actualidade

Em 1992, Simonetta Montemagni e Lucy Vanderwende concentraram-se na extracção de relações a partir da *differentia* e procuraram comparar a utilização de padrões baseados em texto (*string patterns*) com a utilização de padrões baseados na estrutura sintáctica das definições (*structural patterns*) para a construção de enquadramentos semânticos baseados nas definições. Enquanto que os primeiros se limitam a usar determinadas construções de texto específicas das definições como as utilizadas por Chodorow, Markowitz ou Alshawi [Als89], os segundos tomam em atenção as estrutura das árvores sintácticas das frases. A conclusão a que chegam é que os padrões baseados em texto poderão até ser mais fáceis de encontrar e até funcionam muito bem para identificar o *genus* (e assim extrair a relação de hiponímia). Já no que toca a extrair a *differentia*, esse tipo de padrões tem bastantes limitações que só podem ser ultrapassadas com a utilização de padrões estruturais. Os padrões baseados em texto não funcionam bem quando existe um encadeamento de conceitos ao mesmo nível (... *to make* ***laws, rules*** *or* ***decisions***...), quando existem parêntesis no meio da definição, quando é necessário identificar argumentos funcionais ou quando existem relações mais específicas dentro da definição (em *pianta erbacea com* ***bacche di color arancio***, a cor diz respeito às bagas da planta e não à planta.). Apesar do vocabulário presente num dicionário ser mais simples e restrito, ao se utilizar uma gramática

geral para uma língua consegue-se obter informação semântica muito rica sem se estar dependente de especificidades no vocabulário utilizado. Mais recentemente, O'Hara [O'H05] trabalhou no mesmo sentido mas com conceitos e preocupações de uma semântica computacional do século XXI.

Apesar da quantidade de trabalhos explorar a utilização de dicionários para a extracção de conhecimento até à altura, só na década de 1990, depois de várias publicações nesse sentido [WDR93, Van94, Dol94, Van95], a equipa de processamento de linguagem natural de Microsoft criou um recurso completamente independente de um dicionário, o MindNet [RDV98], com base na comparação e processamento de vários dicionários para o inglês. Para a construção do MindNet foi utilizado o analisador sintáctico MEG, utilizado na verificação gramatical do *Microsoft Word 97*. Este analisador produz árvores sintácticas e formas lógicas sobre as quais são aplicadas regras para a extracção de relações semânticas.

3 Agradecimentos

Este relatório foi escrito no âmbito da Linguateca, financiada pela Fundação para a Ciência e Tecnologia e pela União Europeia através dos projectos POSI/PLP/43931/2001 e POSC 339/1.3/C/NAC.

Agradecemos ainda ao Núcleo de Investigação e Desenvolvimento da Porto Editora.

Apesar de não ter validado esta versão final, Nuno Seco participou inicialmente na escrita deste relatório.

Referências

[Als87] Hiyan Alshawi. Processing dictionary definitions with phrasal pattern hierarchies. *Comput. Linguist.*, 13(3-4):195–202, 1987.

[Als89] H. Alshawi. Analysing the dictionary definitions. *Computational lexicography for natural language processing*, pages 153–169, 1989.

[Ams80] Robert Alfred Amsler. *The structure of the Merriam-Webster Pocket dictionary*. PhD thesis, The University of Texas at Austin, 1980.

[Ams81] Robert A. Amsler. A taxonomy for English nouns and verbs. In *Proceedings of the 19th annual meeting on Association for Computational Linguistics*, pages 133–138, Morristown, NJ, USA, 1981. Association for Computational Linguistics.

[Cal77] Nicoletta Calzolari. An empirical approach to circularity in dictionary definitions. In *Cahiers de Lexicologie*, pages 118–128, 1977.

[Cal82] Nicoletta Calzolari. Towards the organization of lexical definitions on a database structure. In *Proceedings of the 9th conference on Computational linguistics*, pages 61–64, , Czechoslovakia, 1982. Academia Praha.

[Cal84] Nicoletta Calzolari. Detecting patterns in a lexical data base. In *Proceedings of the 22nd annual meeting on Association for Computational Linguistics*, pages 170–173, Morristown, NJ, USA, 1984. Association for Computational Linguistics.

[Dol94] William B. Dolan. Word sense ambiguation: clustering related senses. In *Proceedings of the 15th conference on Computational linguistics*, pages 712–716, Morristown, NJ, USA, 1994. Association for Computational Linguistics.

[GGS07] Hugo Gonçalo Oliveira, Paulo Gomes, and Diana Santos. PAPEL - trabalho relacionado e relações semânticas em recursos semelhantes, Dezembro 2007.

[MAE86] Judith Markowitz, Thomas Ahlswede, and Martha Evens. Semantically significant patterns in dictionary definitions. In *Proceedings of the 24th annual meeting on Association for Computational Linguistics*, pages 112–119, Morristown, NJ, USA, 1986. Association for Computational Linguistics.

[MMN80] A. Michiels, J. Mullenders, and J. Noël. Exploiting a large data base by Longman. In *Proceedings of the 8th conference on Computational linguistics*, pages 374–382, Morristown, NJ, USA, 1980. Association for Computational Linguistics.

[O'H05] Thomas Paul O'Hara. *Empirical Acquisition of Conceptual Distinctions via Dictionary Definitions*. PhD thesis, NMSU CS, August 2005.

[RDV98] Stephen D. Richardson, William B. Dolan, and Lucy Vanderwende. Mindnet: Acquiring and structuring semantic information from text. In *COLING-ACL*, pages 1098–1102, 1998.

[Van94] Lucy Vanderwende. Algorithm for automatic interpretation of noun sequences. In *Proceedings of the 15th conference on Computational linguistics*, pages 782–788, Morristown, NJ, USA, 1994. Association for Computational Linguistics.

[Van95] Lucy Vanderwende. Ambiguity in the acquisition of lexical information. In *Proceedings of the AAAI 1995 Spring Symposium Series*, pages 174–179, 1995. Symposium on representation and acquisition of lexical knowledge.

[WDR93] Lucy Vanderwende William Dolan and Stephen D. Richardson. Automatically deriving structured knowledge bases from online dictionaries. In *PACLING 93, Pacific Assoc. for Computational Linguistics*, pages 5–14, 1993.

Tabela de Revisões

Versão	Quem	O quê	Data
0.1	Hugo Oliveira	Primeira versão do documento	29-01-2008
0.2	Hugo Oliveira	Alterações na descrição do PEN e nos Agradecimentos. Escrita da secção Caminho futuro	31-01-2008
0.3	Hugo Oliveira	Correcção da data na capa e pequenas correcções sugeridas pela Diana: Figuras 3 e 4, eliminação das secções 5.3 e Caminho futuro	03-02-2008
0.4	Hugo Oliveira	Eliminação de algumas secções com exemplos mais detalhados da relação Causa.	15-05-2008
0.5	Hugo Oliveira	Eliminação das secções não relativas ao estado da arte.	15-05-2008
1.0	Diana Santos	Pequenas mudanças à estrutura	18-08-2008