Compiling large language resources using lexical similarity metrics for domain taxonomy learning

Ronny Melz*, Pum-Mo Ryu‡, Key-Sun Choi‡

* Institute of Computer Science, NLP Dept., University of Leipzig,
rmelz@informatik.uni-leipzig.de

‡ Computer Science Division, KAIST, KORTERM/BOLA
pmryu@world.kaist.ac.kr, kschoi@cs.kaist.ac.kr

Abstract

In this contribution we present a new methodology to compile large language resources for domain-specific taxonomy learning. We describe the necessary stages to deal with the rich morphology of an agglutinative language, i.e. Korean, and point out a second order machine learning algorithm to unveil term similarity from a given raw text corpus. The language resource compilation described is part of a fully automatic top-down approach to construct taxonomies, without involving the human efforts which are usually required.

1. Introduction

An Ontology, according to its recently extended meaning, is a digital resource that represents shared conceptualizations for a specific domain in an application system. Ontologies proved to be particularly beneficial in application areas such as intelligent information integration, information broking, or natural language processing. Nevertheless the undoubtedly desired broadband usage is inhibited yet by the still time-consuming and cost inefficient construction processes involved.

A taxonomy is a particular domain ontology: It hierarchically relates domain specific terms to each other with increasing specificity from the distinguished root node down to the leave nodes by various kind of relations. It underlies the assumption that domain specific terms are linguistic realizations of domain specific concepts.

The goal of taxonomy learning is to automatically build an ontology of a specific domain and to represent concepts and semantic relations within that domain. We are currently exploring a new building method for specific domain taxonomies benefitting from term specificity and term similarity measures.

Through this contribution we want to publish our methodology of compiling an adequate language resource, which posed a major challenge in our taxonomy learning process. In particular, much attention has to be dedicated to complex Korean derivational morphology and noun composition across the boundaries of word units (eojeols) to achieve profitable results.

The remainder of this contribution is structured as follows: The next section relates our work to previous research. Subsequently, section 3 outlines the overall learning process and section 4 explains the steps to compile the language resource. Section 5 describes the results and draws preliminary conclusions based on a first empirical analysis of the results and points out ongoing investigation. Section 6 closes with acknowledgements and appendix section 7 briefly sketches our data sources.

2. Related Work

Much research of the past addressed automatic taxonomy learning, mainly following three paradigms: lexico-syntactic pattern matching approaches [Hearst '92], learning of vertical term relations [Velardi & al. '01] and lastly statistical approaches based on a term's context distribution.

Significant results were obtained by [Caraballo '99] who examined clustering methods, [Alfonseca & Manandhar '02] focussing on ontology exension, [Cimiano & al. '05] who proposed a vector-space based Formal Concept Analysis and [Yamamoto & al. '05], calculating inclusion relations from word appearance patterns.

While pattern based and vertical relation approaches yield a high precision, they suffer simultaneosly from a poor recall, particularly because patterns are rarely applied in real documents. Likewise, the high-recall distributional approaches suffer from low precision. One problem is that many unrelated terms might co-occur if just occurring frequently enough. Secondly, data sparseness arises as many domain terms are multi-word terms which tend to rarely appear in corpora, hampering the collection of statistically evaluable context information.

All of the existing methods rely on a single metric to learn a taxonomy. For example, [Caraballo '99] used conjunctions of nouns and appositions, [Alfonseca & Manandhar '02], [Cimiano & al. '05] and [Yamamoto & al. '05] used syntactic relations such as modifier-term, verb-subject-term or verb-object-term relation. The strength of our approach lies in its ability to distinguish between specificity and similarity and thus to allow for separately optimized metrics on contextual and term-inherent features.

3. Methodology

In this section we confine the general problem and present the key ideas of the algorithm. We outline the involved steps and motivate our approach by relating it to recently discovered semantic growth principles in human cognition.

3.1. Formulation and restriction of the problem

Ontologies represent shared conceptualizations for a domain. We assume for tractability, that domain specific terms are the directly observable linguistic realizations of these concepts. Given the concepts of a domain, the subsequent task is to interlink these by appropriate relations, giving rise to a graph consisting of a set of

concept nodes and a set of relations labeled from an unproductive, very finite set of labels. More formally, this structure could also be treated as a bipartite graph.

We are particularly interested in taxonomies, a restricted type of ontology. Taxonomic relations typically include *IS-A*, *PART-OF*, *INSTANCE-OF* and various other broader-narrower relations, which constrains the general ontology graph to a tree. As the tree's edges have indeed a natural orientation (increasing specificity in the direction of the leave nodes with respect to the domain under examination), the tree is *rooted*, which means that it has a distinguished root node, i.e. the most general concept characterizing the domain. Relations introducing cycles to the graph (such as *synonymy* and *homonymy*) are not considered in our construction process but could be added in a succeeding step.

We confined the problem further to only account for finding *hypernymy* relations among concepts, which implies trivially the inverse *hyponymy* relation on the same graph. This restriction limits the computational costs and was picked as a typical example for demonstration of the technique. Different relations would obviously require different specificity measures than the measures we investigated, but the general framework extends to those relations equally well.

3.2. Outline of the fundamental process

The taxonomy building process is modeled recursively as a sequential insertion of new terms to the incrementally growing taxonomy. The initial condition is one single state, the root node of the domain. Terms to be inserted are best imagined as a sequence sorted by increasing term specificity. Along with the repeated insertion of terms, the current model reflects increasingly better the rich taxonomic structure as sketched below in Fig. 1.

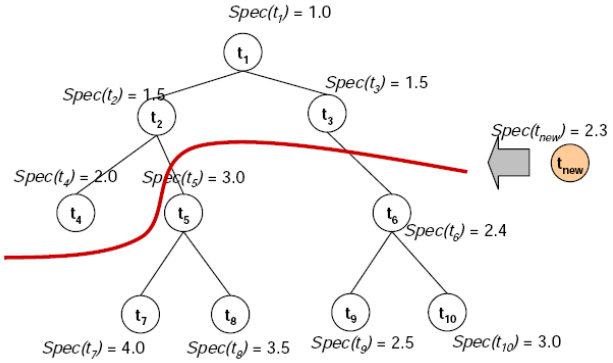

Figure 1 - sketch of the taxonomy growing process

The basic principle of the fundamental algorithm is very simple, as outlined above – now we examine more in detail the implicit prerequisites and the different process stages.

The attachment process requires a concept, or, equivalently, a term T_i to be characterizable by two numerical features. The degree of specificity is reflected by a scalar value, whereas specificity is henceforward assumed to mean always specificity with respect to the domain under consideration. The second numerical feature is complex and reflects similarity to other terms. It can be conceived as a finite set of scalar values which represent the similarity between T_i and other terms T_j of the domain.

Research efforts of the past yielded noteworthy results in specificity [Caraballo & Charniak '99; Sanderson '99; Ryu & Choi '04] as well as similarity measurements [Grefenstette '92; Terra & Clarke '03].

To demonstrate practicability on a prototypic example, we chose the the vast IT field as reference domain. Our starting point was a digitized raw text corpus of several heterogeneous sources from within that domain. The following listing traces the process stages from this starting point up to the domain taxonomy:

1. data cleaning and preparation
2. morphological analysis and POS tagging
3. lexical feature extraction
4. co-occurrence analysis
5. term similarity calculation
6. term specificity calculation
7. actual taxonomy growth

Listing 1 – major process stages

The first step is the irksome obligation faced by everybody dealing with real-world noisy data.

The first three steps together account for the morphologically rich Korean language, provide the lexical features of the text and a set of complex terms actually used in the domain. These term – or a more refined subset of them – will constitute the lexical items of the domain, which in this paper had been assumed given so far.

The following three steps unveil structural information from the corpus exploiting contextual phenomena and large-scale text statistics. It is important to note, that after the seventh step we already compiled a new language resource which differs in significant aspects from the original raw corpus: we extracted the atomic concepts and enriched them with structural information.

This contribution aims at describing the construction of the language resource up to the fifth step of the preceding process outline. For the detailed term specificity calculation refer to [Ryu & Choi '04]. The final steps and results will be covered in future publications.

3.3. Motivation and justification of the approach

The conception of the large-scale structure of real world networks has changed dramatically in the very recent past due to the availability of detailed information about large and complex real world graphs, less expensive computational power and an immense worldwide interest. This rise in interest is explainable from the insight, that structure implies the dynamical properties of a graph, e.g. the propagation of information on the internet, the spread of jokes or pandemics in social networks or the distribution in electrical power networks [Wang & Chen '03].

Semantic networks such as WordNet [Fellbaum '98], Roget's Thesaurus [Roget '11], and free naïve associations [Nelson & al. '99] have been found to display essentially the same properties as their counterparts in IT or the social sciences [Steyvers & Tenenbaum '05].

The most successful model in explaining these properties is that of [Barabasi & Albert '99]. It explains the development of structure through a dynamical process driven by *incremental growth* and *preferential attachment*. They propose an algorithm, in which one new node per time instance is attached to the network (incremental growth). This new node is more likely to be linked to the

fewer important nodes[1] of the already existing network, than to the abundant less important ones (preferential attachment). Node importance corresponds to the (in)degree of a node and the attachment probability of a new node is made directly proportional to the importance. This gives amazing insights in the growth of the link structure of internet pages and it is the same insight, that has led the most popular search engine algorithms to great commercial success.

For the growth of semantic networks, different features are prevalent. "Following the suggestions of many researchers in language and conceptual development, [Steyvers & Tenenbaum] assume, that semantic structures grow primarily through a process of differentiation: the meaning of a new word or concept typically consists of some kind of variation on the meaning of an existing word or concept."

Unlike the above quoted Steyvers and Tenenbaum, our interest is not to develop an abstract model of semantic growth. Our aim is to extract the structural information latently contained in the raw text and to make this knowledge explicit in form of domain taxonomies, i.e. to reverse-engineer the structure predicted by the probabilistic model of semantic growth. Hence, there are no random variables involved in our attachment process. Moreover, as pointed out in the previous sections, we focus on a particular kind of semantic differentiation: specification. New nodes are attached in function of their specificity and conceptual relatedness to pre-existing nodes.

Although having severe limitations as a general model of semantic structure due to its strong constraints (acyclicity precludes synonymy relations, e.g.), a tree-structured hierarchy nonetheless provides a particularly economical system for taxonomic knowledge, as has been shown by [Collins & Quillian '69].

4. Implementation

This section descibes the setup of the building process and covers steps one through five of listing 1.

4.1. Data cleaning and preparation

The heterogeneity of the data sources as well as the sheer mass of data required extensive cleaning before starting the following, time complex operations. This cleaning took place partially before and partially after tagging: before tagging, sentence boundaries had to be detected and aligned and text layout elements, e.g., to be removed. After tagging we could reliably detect and remove fragments of formulae from the patent documents or map numbers and numerals to a single token.

4.2. Morphological analysis and POS tagging

An agglutinative language as morphologically rich and complex as Korean requires morphological tagging of sub-eojeol components before a statistical text analysis.

On the one hand, compound nouns frequently span over several eojeols (sentence units). Since compound nouns are very likely potential candidates to belong to the domain terminology, we have a strong interest to extract them with high recall (see next subsection for details).

1 since Barabási and Albert do not restrict the graph to be acyclic, a new node may be linked to multiple pre-existing nodes

On the other hand, to consider words only differing in their attached case particles as distinct terms would not be of great benefit for the applicability of the taxonomy and would in addition increase data sparseness. Many glutei do not reflect semantic variety, which justifies their removal (stemming).

Our strategy is as follows: we strip each of the particles (nine case particles [jc*] and two auxiliary particles [jx*]) from each eojeol. The only exception to that rule poses the predicative particle (copula) [jp], because it actually modifies the semantics of the eojeol being attached to. Beyond that, we also strip potential endings [e*] and symbols such as brackets, commas or quotation marks from eojeols. Other parts of speech such as predicates, modifiers or affixes remain unchanged. A source sentence would thus be mapped to a simplified target sentence where supplemental agglutinative information is largely removed, cf. the following example:

ncpa+jcm	**nnc+nbu+jcs**	**maj**	**pvg+ep+ef**	**sf**
격동+의	한+해+가	또	저물+었+다	.
격동	한해	또	저물	.
turbulent	year	again	came to close	

We use a tagger for Korean, which is described along with its tagset in [Lee, Choi, Kim '93; Lee, Cha, Lee '02].

4.3. Lexical feature extraction

The extraction of relevant lexical features plays a key role, since the similarity measurement will be based upon these features. Concluding from previous research (cf. [Ryu & Choi '05]), especially compound nouns carry a signifcant amount of domain knowledge. Moreover, their specificity tends to rise with increasing complexity: 'network' or 'computer' as singularly occurring concepts may very well be considered general language terms, whereas 'computer network' or 'local area network' are quite characteristic concepts of the IT domain, e.g. Thus, it is crucial to identify and extract compound nouns.

Given the POS-tagged corpus, we extracted compounds by the following incremental process: add to a single noun also the subsequent component if it is tagged

- common noun [nc*]
- foreign character sequence [f]
- dash [sd] or noun derivative suffix [xsn] only if a further noun follows.

Eojeols terminating in affixes having other POS tags – especially case particles [jc*]– are not valid internal parts of a compound noun. They terminate the chain contruction process. Notably, proper nouns [nq] too are not considered valid components for compounding. From the sentence

보/pvg + ㄴ/etm	this
고안/ncpa + 은/jxc	design
전원스위치/ncn + 에/jca	electric switch
관하/pv + ㄴ/etm	about
것/nbn + 으로서/jca + ,/sp	(something)
특히/mag	particularly
전기/ncn	electric
전자기기/ncn	electonic device
본체/ncn + 외부/ncn + 에/jca	on external body

설치되/pv + 어/ecx + 지/px + 는/etm be installed
스위치놉/ncn + 이/jcs switch knob
외관/ncn + 상/xsn appearance
미려하/pa + 게/ecx + 하/px + 고/ecc elegant
내부/ncn internal
스위치/ncn switch
접점/ncn + 이/jcs contact point
완벽하/pa + 게/ecs perfectly
이루어지/pv + ㄹ/etm + 수/nbn be established
있/pa + 게/ecs + 하/pv + ㄴ/etm can
것/nbn + 에/jca (something)
주안점/ncn + 을/jco + 두/px + ㄴ/etm focused
것/nbn + 이/jp + 다/ef (something)
./sf

the extracted set of compound nouns {'고안', '전원스위치', '전기 전자기기 본체 외부', '스위치놉', '외관', '내부 스위치 접점', '주안점'} were added to the compound noun list. (The sentence roughly means "The design concerns electric switches, particularly focused on making switch knobs, installed on the external body of electric or electronic devices, looking elegantly and establishing a perfect contact point with the internal switch")

This list of 8'551'598 (compound) nouns constitutes the biggest fraction of the vocabulary of 14'918'957 word types in total after merging it with the remainder of the lexical features. We do not yet see the necessity for an exhausting analysis of the validity of the extracted compounds, as mother tongue speakers verified the result satisfactory.

In a first order approximation, these nouns (some of which are compounds) already are the terminology of the domain. This is a very coarse grained criterium, trading off precision for high recall. A simple difference analysis contrasting the domain and general noun distribution would quickly narrow down the extracted terms to a fairly smaller subset. Many sophisticated terminology extraction methods –linguistical, statistical or hybrid– have been intensely studied. Some references can be found in [Oh, Lee & Choi '00; Witschel '05].

We postponed this pruning as it involves irreversible information loss. Indeed, the refinement to a subset can be applied at any future stage, e.g. immediately before step seven (the actual taxonomy induction).

Furthermore it is essential to conserve the rest of the vocabulary as well, consisting mainly of open class POS such as verbs or adjectives because our similarity calculation is based on a second order approach. Their purpose is not to be included in the taxonomy, but to serve as features to characterize the actual terms.

4.4. Co-occurrence analysis

To measure similarity between words, we start by gathering their co-occurrence statistics from the text base. We use the log likelihood test proposed by [Dunning '93]. In brief, this corresponds to an independence test of two binomial distributions, i.e. whether word w_A occurs (associated to event A) independently of word w_B or not.

$$\frac{P_{Bi(n_{AB};\,n_B)}(A)\cdot P_{Bi(n_A-n_{AB};\,n-n_B)}(A)}{P_{Bi(n_{AB};\,n_B)}(A|B)\cdot P_{Bi(n_A-n_{AB};\,n-n_B)}(A|\neg B)}$$

Formula 1 - the likelihood ratio formula

The reverse of this quantity (or its negative logarithm, due to monotony) serves to rate statistical dependence, which is interpreted as significance measure of word co-occurrence.

We used the ConceptComposer by [Schmidt '00] to compute our results. The language independence had already been shown by [Biemann et al. '04].

Since we chose a vocabulary of the size of almost 15 million items, we theoretically had to cope with a set of co-occurrences of quadratic size, more than twelve orders of magnitude. Fortunately, inter alia due to finite sentence length, the co-occurrence matrix remains sparse. The ConceptComposer uses a reverse indexing (sentence for wordform type) and pruning techniques to efficiently sum up the counts required in the above formula. Since this method only considers pairs actually found in the data, it scales well to even large amounts of input. The size of co-occurrences tends to increase rather linear: we achieved about 200 million co-occurrence relations out of 23 million sentences.

4.5. Term similarity calculation

In the next step, we compared the contexts of each of the terms. Here, context means a set of lexical features w_B stored for each term w_A. The set only contains the most significant co-occurring words along with their significance values as measured by formula 1. This set can be conceived as the the typical, global syntagmatic context of term w_A, condensed from all the local contexts in the input data.

Since the iteration over the term pairs is computationally more costly than similarity calculation itself, we chose different formulae in parallel to measure similarity, among them *Count*, *Cos*, *Dice* and *Jaccard*. Count is simply the number of identical items in two context sets. The dice measure doubles the count and divides by the summed cardinality of the two individual context sets. Jaccard divides count by the cardinality of the union of the two individual context sets. For cosine similarity calculation, the context sets of w_A and w_B are projected into a real vector space $\mathbb{R}^N$, where N is the cardinality of the union of the context sets, the coordinates take the significance values of lexical items in the context sets. Now the cosine similarity is defined as the dot product of the two vectors after having them normalized to unit length.

Unfortunately, the similarity calculation is bounded below by quadratic time complexity in the size of the term list. Although the task posed a heavy computational burden, we did not face scalability problems.

These relations calculated are of second order in the sense that we used association measures to discover relations of first order in the text and subsequently use them to calculate similarity among terms. The result were around 370 million term pairs.

Saussure distinguished in his noteworthy 'Cours de Linguistique Générale' [Saussure '16] two fundamental kinds of relationship between signs: '*syntagmatic*' and '*paradigmatic*'. In brief, two signs relate syntagmatically, if they complement or support each other, both functionally and in content. They relate paradigmatically, if the two signs share substantial characteristics, again both functionally and in content (cf. also [Rapp '02]).

For example, in 'the sun is shining bright', 'bright' is an adverb to 'shining', or 'shining' a verb complementing 'sun'. Both are syntagmatic relations. 'Sun' shares important characteristics with other sources of light such as a 'candle', which could be substituted for sun in this context. 'Sun' and 'candle' are related paradigmatically.

The co-occurrence formula filters the most significant lexical features out of all the given contexts, hence we can motivate the extracted sets as statistically significant syntagmatic relations. The similarity relations based on these syntagmatic relations can thus be conceived as paradigmatic.

5. Results

Subsuming, we constructed a structured language resource derived from raw text material and confirmed the language independency and scalability of our methods up to very high numbers of input tokens.

Since we calculated similarity with respect to different metrics, we are now free to choose the most adequate for taxonomy construction. We do not have the capacities to base an evaluation on extense human examination and direct evaluation is not possible due to the absence of a gold standard. We propose the following way to automatically determine the best metric.

[Budanitsky & Hirst '06] evaluated five WordNet-based similarity metrics by comparing them directly to human ratings. Although each of the metrics agreed similarly well with the human ratings, [Leacock and Chodorow '98]'s method, sim_{LC}, showed the highest correlation coefficient. Leacock and Chodorow proposed the following formula for computing the scaled semantic similarity between two words in WordNet:

$$sim_{LC}(w_1, w_2) = -\log \frac{len(w_1, w_2)}{2 \cdot max_{w \in WordNet} depth(w)}$$

For a given thesaurus the denominator signifies twice its maximum depth and len(w1, w2) the length of the shortest path between two words.

We make use of this finding and relate it in a second step to the similarity values obtained by us. We translated the IEE's Inspec thesaurus (scientific and technical information service) and compared 3'768 term pairs rated by sim_{LC} to tantamount pairs from our corpus in a spot check experiment. The correlation remained below our expectation, which is partially due to the inaccurracies of the similarity calculation, partially due to translation

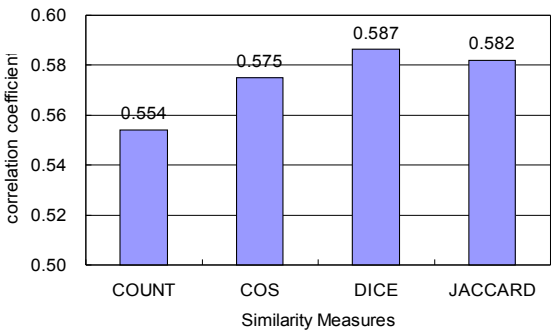

Figure 2 -- correlation between sim_{LC} and our obtained results

inaccurracies. Nonetheless we could discriminate the dice coefficient to correspond best to human intuition without directly involving costly human experiments. For an extensive evaluation of lexical acquisition see also [Bordag & al. '05]

Currently, we are setting up the final steps for taxonomy learning: specificity measurement, a reduction of the terminology to get a more concise, high-precision set of terms, and lastly the recursive learning function.

6. Acknowledgements

In this section, we like to thank for the extense conceptual support of our supervisors. We are also greatful to text:tech GmbH for providing us with the ConceptComposer suite and Stefan Bordag for his support and his implementation of the similarity module. Thanks go to the anonymous reviewers as well, who constructively criticized the preliminary version of this paper. Finally, we want to thank the German Academic Exchange Service (DAAD), without whose financial support this work would not have been possible.

7. Appendix – Data Sources

A huge amount of IT related documents in Korean has been collected, a total of 783157 documents mixed from different sources.

- The largest amount cover Korean IT patents and registered patterns in the period from 1971 until 2004 (720140 documents), provided by the Korean Patent Agency. To include a broader range of terminology, further 63017 documents, partially significantly longer than the patent documents, have been included:
- '전자신문', the daily published "Electronic Times" (www.etnews.co.kr) from 1994 to 2004
- Korean IT laws by the Korean Supreme Prosecutors' Office (http://icic.sppo.go.kr/; related to internet crime) and by the Korean Ministry of Information and Communication (http://www.mic.go.kr; IT related strategies)
- conference papers in IT (Korean Information Science Society, Korean Information Processing Society and HCI Conference) from 2002 to 2004
- 123 recently published text books on IT, crossing all levels of education – from elementary school up to university level.

8. References

[Barabási & Albert '99] Barabási , A.L., & Albert, R. (1999). Emergence of scaling in random networks. Science, 286 , 509-512.

[Alfonseca & Manandhar '02] Alfonseca, E., Manandhar, S. (2002) Extending a Lexical Ontology by a Combination of Distributional Semantics Signatures. In proceedings of 13th International Conference, EKAW.

[Biemann et al. '04] C. Biemann, S. Bordag, G. Heyer, U, Quasthoff, C.Wolff (2004), Language-independent Methods for Compiling Monolingual Lexical Data, Proceedings of CICLing; Springer LNCS 2945, pp. 215-228

[Biemann, Shin & Choi '04] Biemann, C., S.-I. Shin, und K.-S. Choi: Semiautomatic extension of corenet using a bootstrapping mechanism on corpus-based co-occurrences. In: Proceedings of the 20th International

Conference on Computational Linguistics, COLING04, Genf, Switzerland, 2004.

[Bordag & al. '05] S. Bordag, H.F. Witschel, and T. Wittig. Evaluation of Lexical Acquisition Algorithms. In B. Fisseni, H.-C. Schmitz, B. Schröder, and P. Wagner, editors, Proc. of GLDV-Tagung 2005, pages 449-461, Frankfurt, 2005. Peter Lang.

[Budanitsky & Hirst '06] A. Budanitsky, G. Hirst (2006), Evaluating WordNet-based Measures of Lexical Semantic Relatedness, Computational Linguistics, Vol. 32, Num. 1, 2006

[Caraballo '99] Caraballo, S. (1999) Automatic construction of a hypernym-labeled noun hierarchy from text. In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 120-126

[Caraballo & Charniak '99] S. A. Caraballo and E. Charniak (1999), Determining the Specificity of Nouns from Text, In the proceedings of the Joint SIGDAT Conference on EMNLP and Very Large Corpora

[Cimiano & al. '05] Cimiano, P., Hotho, A., Staab, S. (2005) Learning Concept Hierarchies from Text Corpora using Formal Concept Analysis. Journal of AI Research, Vol. 24, pp. 305-339

[Collins & Quillian '69] Collins, A.M., & Quillian, M.R. (1969). Retrieval time from semantic memory. Journal of Verbal Learning and Verbal Behavior, 8 , 240-248.

[Dunning '93] Ted E. Dunning (1993). Accurate methods for the statistics of surprise and coincidence. Computational Linguistics, 19(1):61–74.

[Fellbaum '98] Fellbaum, Christiane. 1998. A semantic network of english: The mother of all wordnets. Computers and the Humanities, 32:209–220.

[Grefenstette '92] Gregory Grefenstette (1992), “Finding semantic similarity in raw text: the deese antonyms”. In Robert Goldman, Peter Norvig, Eugene Charniak, and Bill Gale, editors, Working Notes of the AAAI Full Symposium on Probabilistic Approaches to Natural Language, pages 61–65, Menlo Park, CA. AAAI Press.

[Hearst '92] Hearst, M. (1992) Automatic Acquisition of Hyponyms from Large Text Corpora. In Proceedings of the 14th International Conference on Computational Linguistics

[Leacock and Chodrow '98] Leacock, C. and Chodrow, M. (1998), Combining local context and WordNet similarity for word sense identification, In WordNet: An electronic lexical database, MIT Press

[Lee, Cha, Lee '02] Gary Geunbae Lee, Jeongwon Cha, Jong-Hyeok Lee. Syllable pattern-based unknown morpheme segmentation and estimation for hybrid part-of-speech tagging of Korean. Computational Linguistics, Vol 28, No 1, pp 53-70, March 2002

[Lee, Choi, Kim '93] W.J. Lee, K. S. Choi, K. C. Kim (1993). An Automatic Tagging System for Korean Texts. Proceedings of the spring conference of Korea Information Science Society, pp. 805-808

[Nelson & al. '99] Nelson, D.L., McEvoy, C.L., & Schreiber, T.A. (1999). The University of South Florida word association norms.

[Oh, Lee & Choi '00] Jong-Hoon Oh, Kyung-Soon Lee and Key-Sun Choi. 2000. Term Recognition Using Technical Dictionary Hierarchy. In Proceeding of the 38th Annual Meeting of the Association for Computational Linguistics, ACL 2000

[Roget '11] Roget, P.M. (1911). Roget’s Thesaurus of English Words and Phrases (1911 edition) . Available from Project Gutenberg, Illinois Benedictine College, Lisle, IL.

[Rapp '02] Reinhard Rapp (2002), The Computation of Word Associations: Comparing Syntagmatic and Paradigmatic Approaches

[Ryu & Choi '04] Pum-Mo Ryu, Key-Sun Choi (2004), "Determining the Specificity of Terms based on Information Theoretic Measures" , Proceedings of CompuTerm 2004, 3rd International Workshop on Computational Terminology, Coling 2004, pp. 87-90

[Ryu & Choi '05] Pum-Mo Ryu, Key-Sun Choi (2005), "An Information-Theoretic Approach to Taxonomy Extraction for Ontology Learning" , In P. Buitelaar, P. Cimiano and B. Magnini(eds.), Ontology Learning from Text: Methods, Evaluation and Applications, Vol. 123, Frontiers in Artificial Intelligence and Applications, IOS Press, Amsterdam, July 2005

[Sanderson '99] Mark Sanderson (1999), Deriving concept hierarchies from text, In the proceedings of the 22th Annual ACM S1GIR Conference on Research and Development in Information Retrieval

[Saussure '16] Ferdinand de Saussure (1916), Cours de Linguistique Générale, Payot, Paris

[Schmidt '00] Fabian Schmidt, 2000: Automatische Ermittlung semantischer Zusammenhänge lexikalischer Einheiten und deren graphische Darstellung, Diplomarbeit, Universität Leipzig

[Steyvers & Tenenbaum '01] Mark Steyvers and Joshua B Tenenbaum (2005), The large-scale structure of semantic networks: statistical analyses and a model for semantic growth, Cognitive Science, 29(1)

[Terra & Clarke '03] Terra, Egidio and C. L. A. Clarke (2003). Frequency estimates for statistical word similarity measures. In HLT-NAACL 2003, pages 165–172.

[Wang & Chen '03] Xiaofan Wang and Guanrong Chen, Complex Networks: Small-world, scale-free and beyond, IEEE Circuits and Systems Magazine, Vol. 3, No. 1, 2003, 6-20.

[Watts & Strogatz '98] Watts, D.J., & Strogatz, S.H. (1998). Collective dynamics of ‘small-world’ networks. Nature, 393, 440-442.

[Velardi & al. '01] Velardi, P., Fabriani, P., and Missikoff, M. (2001) Using Text Processing Techniques to Automatically enrich a Domain Ontology. In Proceedings of the ACM International Conference on Formal Ontology in Information Systems

[Witschel '05] Witschel, Hans Friedrich “Terminologie-Extraktion - Möglichkeiten der Kombination statistischer und musterbasierter Verfahren” In der Reihe: "Content and Communication - Terminology, Language Resources and Semantic Interoperability". Würzburg: Ergon Verlag, 2004.

[Yamamoto '05] Yamamoto, E., Kanzaki, K. and Isahara, H. (2005) Extraction of Hierarchies Based on Inclusion of Co-occurring Words with Frequency Information. Proceedings of 9th International Joint Conference on Artificial Intelligence, pp. 1160-1167