IVal – An Alternative WordNet Browser for Evaluating Semantic Informativeness of Concepts

Jürgen Reischer
Information Science
University of Regensburg
93040 Regensburg
juergen.reischer@sprachlit.uni-regensburg.de

Abstract

This paper describes theoretical and practical aspects of a procedure for calculating semantic informativeness of concepts on the basis of WordNet. The IVal system is introduced which provides enhanced functionality for accessing the WordNet database including the computation of the concept's informativeness, the decompositional analysis of terms, and an interface for extending the lexicon.

Introduction

The WordNet lexical database contains a huge amount of linguistic information about terms and concepts. The relation between them can be described by the classical Saussurean model of signs [Saussure 1967]: a term (signifier) is associated with a concept (signified) with the concept being the sense of the term (signification). Another relation holds between two concepts, which may be linked by a semantic relation. To exploit the information coded in this lexical system of signs, WordNet provides a browser for accessing these data. In the following sections, a browser with extended functionality is introduced: the IVal system is designed to provide enhanced access to the linguistic data coded explicitly and implicitly in WordNet's lexical database. The system allows extended retrieval and analysis of lexical items and supports the expansion of the lexicon for new items. These components will be described below. One aspect, the calculation of semantic informativeness of concepts, will be examined in more detail.

1 The IVal System

The WordNet system contains a vast amount of linguistic information coded explicitly or implicitly in its lexical database [cf. Fellbaum 1998]. To provide systematic access to these data, a retrieval interface for the user is needed. The original browser offers an easy way to access the linguistic database of WordNet. For enhanced functionality, e.g. to extend the WordNet lexicon, no interface is available at the moment. IVal is designed to fill this gap and to provide further extended access to the WordNet database. In the following subsections some of the components already available in IVal will be described in detail.[1]

1.1 Extended Browser Functions

IVal extends the functionality of the original WordNet browser in several ways. Some general extensions include the output of the term's sense analysis, which may be filtered for lexical categories and types of semantic relations to be displayed. Besides this, every output expression is hypertextified so that the user may follow further interesting items immediately. More specific enhancements will be described below.

1.1.1 Decomposition of Terms

In the WordNet browser, every term entered will be analysed for its associated concepts, i.e. the senses of the term according to the WordNet lexicon. Terms may be inflected or uninflected, single- or multi-word expressions of any content

[1] The IVal system in its actual version is available via 'http://www.lingua-ex-machina.de'.

word category, like ‘account’ (noun) or ‘take into account’ (verb). Terms not in the database cannot be analysed by the WordNet browser due to the missing morphological decomposition. For that purpose, IVal provides a morphological parser which tries to decompose expressions not found in the database.

The decomposition is based on a binary word grammar to structurally analyse derivatives and compounds. Morphological elements are internally annotated with linguistic information so that unplausible decompositions are avoided whereever possible. For example, the suffix ‘-ly’ can neither be attached to a prefix (*‘un-’ + ‘-ly’) nor to a verb (*‘retrieve’ + ‘-ly’). Elements to be combined include affixes and stems as well as combining forms like ‘biblio-’ and ‘-phil’ (note *‘-phil’ + ‘biblio-’).

Furthermore, derivatives and compounds may be combined to complex hyphenated expressions like ‘informativeness-calculation’, which have to be decomposed as well into their partial expressions. The IVal morphological parser employs a hierarchical analysis method to deconstruct any kind of complex expression, as can be verified in the analysis output of the browser.[2]

1.1.2 Calculating Semantic-Thematic Distance

An interesting feature of the WordNet database is the fact that concepts are linked to a conceptual network by several semantic-thematic relations. Especially noun and verb concepts are connected to a hierarchical system which provides implicit information about the semantic or thematic relatedness between concepts. To exploit this information, the IVal browser allows to measure the semantic-conceptual distance between two concepts as linked in the WordNet database.

Basically, two distance measurements are available: First, the *semantic* distance between two concepts A and B can be measured on the basis of hyponymy and hypernymy alone. Second, a *thematic* distance measure including further relations like meronymy, holonymy, or topical relations is provided. Both measures use the same algorithm to compute the minimum distance between two concepts A and B in terms of intermediating concepts from A to B.

The algorithm, which cannot be described in detail here, uses a parallel search strategy in order to minimise search time – especially for thematic relatedness where several types of relations are possible to be followed at every concept. At the moment, there is no weighting of the qualitatively different relation types to simulate the judgements of human raters with respect to thematic distance. This is intended to be improved in the near future.

1.2 Extensions of the Database

The WordNet lexical database is designed as a semantic network of conceptual units. Thus, only content words are encoded in the lexicon excluding the function word categories preposition, pronoun, determiner, conjunction, and auxiliary verbs. However, they may be needed for the analysis of texts, e.g. in order to detect syntactic structures. IVal extends the WordNet database by function word terms of the categories mentioned above.

Although this is already an improvement, it may be the case that one wants to extend the content categories for new terms and concepts, too. This provides not only the opportunity to add missing orthographical variants like ‘online’ (instead of ‘on-line’) but also to complete existing and model new thematic domains (cf. the missing term and concept ‘Microsoft’). Even the construction of an entirely new ontology in the WordNet format would be possible.

For this purpose, IVal provides an interface for the extension of content elements comprising the definition of terms and concepts as well as the semantic relations between them. The interface offers also an enhanced possibility to retrieve concepts by the synonyms expressing it and/or by keywords of its gloss. Additionally, the expanded IVal database, which is an information enriched version of the original WordNet database, can be exported in the WordNet format.

[2] This does not include multi-word terms within hyphenated expressions.

2 Semantic Informativeness

One capability of the IVal browser will be discussed in detail here: the calculation of semantic informativeness of concepts. Informativeness is to be understood as the information content of a concept relative to a conceptual hierarchy as available in WordNet. The theoretical and practical aspects of this approach to the quantification of semantic informativeness will be explicated in detail below. Finally, we will discuss some possible applications and give an outlook of further research in this area.

2.1 Theoretical Aspects

The notion of information content has been used in the context of Shannons mathematical theory of communication [Shannon 1948], later called 'information theory'. Information content in this theory is calculated as the binary logarithm of the reciprocal probability of the occurrence of an event (e.g. the appearance of a sign) measured in bits. Although this quantity is called *information content*, it has nothing to do with semantics. Shannon himself stated that his measure concerns only the technical problem of communication and abstracts away from semantic aspects [Shannon 1948: 1].[3]

Thus, Shannon's measure may only be applied to signifiers in the Saussurean sense without considering their meaning or their impact on an interpreter. In order to turn to semantics and to semantic information content, we have to consider *concepts* in the sense of signifieds. Concepts, then, are to be understood as the senses of signs (e.g. terms, words), as realised in the WordNet lexicon. The informational content of a concept may be interpreted as the defining semantic features of that concept discriminating it from other concepts of the same system. For example, the upper noun concept hierarchy of WordNet splits up immediately into physical and abstract entities with all subconcepts having this semantic feature by inheritance.

The taxonomy as provided by WordNet's conceptual hierarchy abstracts away from the specific use of a concept in a certain (con)text by a certain user (i.e. writer/reader or speaker/hearer). In this sense, the informativeness of a concept as defined in WordNet must be regarded as an *objective* measure of the information contained in a concept common to all contexts and users. If there is no common core of semantic features of a concept in all its uses, communication between individuals in different contexts would be impossible. Thus, a WordNet concept represents a *type* with a certain informational potential which is actualised (set free and enriched) by its usage as a token.

In the following section, we are interested in this objective information content as provided by the WordNet taxonomy. This does not deny subjective informativeness of a concept when it is interpreted by an individual. However, the semantic content of a concept grasped by an interpreter cannot be quantified.

2.2 Practical Aspects

A conceptual hierarchy as it is realised in WordNet is based on the principle of inheritance: The most general concept is placed at the top node of the taxonomy (e.g. 'entity'), where more specific concepts are attached below successively. A subconcept inherits the features of its superconcept(s) and adds one or more new features further specifying it. At the bottom of a taxonomy we find *instance* concepts which do not (sub)classify any more but describe real singular entities in the world (e.g. the concept 'Albert Einstein'). This taxonomical structure allows a relative measurement of the informational content of a concept according to its position in the hierarchy.

The calculation of information content consists of three basic steps, which will be explained in detail below: (1) We have to determine the vertical position of a concept C in the hierarchy, i.e. the distance D_R from the top node R to the concept C in question (there may be more than one path from the root R to the concept C due to several possible hypernyms of C, so we have to build an average distance); (2) the depths of all

[3] A theory of semantic information content of statements may be found in [Bar-Hillel & Carnap 1953]. However, this approach is not computerisable.

subtrees of the concept C as opened by its hyponyms have to be calculated and averaged; the subtree depth is the averaged distance D_I from the concept C to all instance nodes I_i of the respective subtree below C; (3) the information content IC of the concept C can now be calculated as the ratio of D_R and D_I, i.e. $IC_C = D_R / D_I$. An example will help to clarify the procedure.

Suppose we have the following small hierarchy with eight concepts:

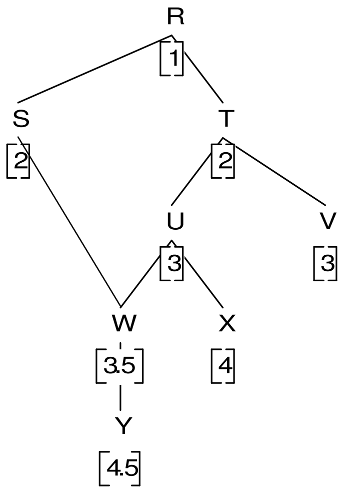

Fig. 1: average vertical concept positions (step 1)

The numbers below the concept identifications represent the averaged distances from the root node R to the respective concepts S to Y.[4] A special case concerns the concepts W and Y: Because two paths with different distances lead to them – R-S-W(-Y) and R-T-U-W(-Y) – we have to average the distance values of 3 and 4 to 3.5 (4 and 5 to 4.5, respectively). This can be justified by the fact that, obviously, W and Y inherit different features from both the S and T paths to W and Y with different semantic content. The absolute value of this added information (e.g. in semantic bits) cannot be measured and must be replaced by the relative values according to the actual hierarchy. In contrast to X, for example, W inherits information from both S and U where U seems to contribute more defining features than S. Thus, the total information gain of W must finally be greater than the gain provided by S alone.

In order to locate the relative vertical position of a concept in the hierarchy, we do not only need the distance to the root node but also the distances to the terminal nodes (instance concepts in this case). The following tree shows the averaged subtree depths:

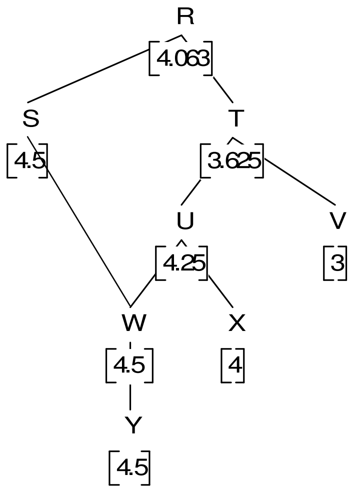

Fig. 2: average subtree depths (step 2)

For terminal nodes, the depth of their ‘subtree’ is identical to their vertical position.[5] For all other nodes, these values must be returned in a recursive procedure. For concepts with multiple subtrees their depth values have to be averaged: e.g. U receives a value of 4 and 4.5 from its two hyponyms resulting in the average 4.25. There is no weighting of the different subtree branches, for example on the basis of the number of subnodes, because in a fully fledged taxonomy all semantic distinctions should be considered equally important. If a concept like T has four subconcepts in branch U and only one in branch V, then both branches make an equally important statement about the semantic specificity T al-

[4] We start counting with level 1 because level 0 is assumed to be the upmost abstract node comprising all kinds of concepts including other lexical categories like verbs and adjectives.

[5] In the above hierarchy, all terminal nodes are assumed to be instances. If this is not the case as in the WordNet hierarchy, virtual instances have to be added again (at least in the noun taxonomy).

ready has and the possible semantic space which can be opened further below T.

In the final step to calculate the semantic information content of a concept, the ratio of its vertical position and its depth(s) is to be calculated. For example, $IC_R = 1 / 4.063 \approx 0.25$, $IC_U = 3 / 4.25 \approx 0.71$, $IC_Y = 4.5 / 4.5 = 1$ (values are normalised). Instances like Y must have an information content of 1, because they provide as much information as possible to identify exactly one single entity in the world. Informativeness 0 is only found in the virtual top node above R, which does not provide any defining or discriminating semantic features at all (cf. footnote 4).

If new items are added to the lexicon permanently or temporarily, all informativeness values of the hierarchy have to be recalculated according to the procedure specified above. This especially holds for compound concepts of the form ZY, where Y is the semantic head of the complex concept. ZY, then, is located below Y one level deeper in the respective hierarchy (see fig. above).

2.3 Discussion and Future Work

The information measure as explicated above is to be understood primarily as a *basic principle* to calculate semantic informativeness of concepts automatically. The details like weighting or level counting of (virtual) roots and instances are certainly debatable. Furthermore, the values gained cannot be better as the organisation of the hierarchy itself: the better the hierarchy with respect to a well-designed taxonomy, the more appropriate calculations will be. An ideal ontology should be, for example, a binary structure on every conceptual level making only basic semantic distinctions (like physical vs. abstract, natural vs. artificial entities etc.), leading finally to a well-formed taxonomy with equally sized subtrees.

The value of informational content represents the maximum amount of information the concept can provide if used in a (con)text by a user. In this sense, it is an information *potential* which can or cannot be exploited by a possible recipient. The quantity of information actually arriving at the interpreter could at least be estimated by an additional factor, e.g. the familiarity of the *concept*. The frequency of a *term* can be regarded as an indicator for its familiarity and therefore readability [cf. Mikk 2000: 79 f]. Analogously, the frequency of a concept (as conveyed by a term) may be used as an indicator of how much information content or potential arrives on average at a normal user and how understandable the concept will be. For example, if a concept with a high information content of 0.9 has an extremely low frequency, the probability that the user is unable to interpret the concept due to its unfamiliarity is very high. Thus, low frequency concepts with high information values may transfer less information on average than very frequent concepts with information values of, say, 0.5 (because *most* interpreters are familiar with it).

Thus, one kind of an informativeness measure considering the subjective factors of interpretability may be the product of semantic information content and familiarity (as indicated by frequency): $IV_C = IC_C \times G(FC_C)$ (with IV as subjective information value of concept C, FC = frequency count of *concept* C as provided by WordNet, and G as an appropriate function). It is a future task to determine G so that $G(FC_C)$ reflects the (normalised) familiarity of C and IV_C is an appropriate measure of subjective semantic information value.

In applications, semantic information content as defined above may be used, for example, to create a profile of the semantic structure of a text indicating more general or more specific passages. The latter may be harder to understand but may contain more potential information. Further, we can think of extracting too general and too specific concepts from a document's term index based on semantic properties (and not just on frequencies). Within the IVal project itself, ICs will be used with the same intention in forthcoming functions like thematic chain construction. An interesting side effect of automatic IC calculation was the detection of hierarchisation errors in WordNet 2.1, when instances are classified under instances (e.g. ‚Paternoster' is conceptualised as an instance of ‚Lord's Prayer' which itself is an instance of ‚prayer').

IVal is work in progress. The final purpose of the system is text analysis with respect to questions of informativeness (e.g. passage retrieval, summarising). For that aim, further components will be realised within the IVal system, e.g. the construction of thematic chains on the basis of lexical chains and thematic distance procedures.

Conclusion

This paper described the IVal system, a WordNet browser with extended functionality like term decomposition or lexicon expansion. One component of the computation system was explained in detail: a basic method for calculating semantic informativeness of concepts which are structured in the conceptual hierarchy of WordNet. The IVal systems extracts the information implicitly coded in the taxonomy and makes it explicit for further use.

References

[Bar-Hillel & Carnap 1953] Bar-Hillel, Y. & Carnap, R. (1953): Semantic Information. In Jackson, W. (Eds.): *Communication Theory*. London: Butterworths Scientific Publications, pp. 503–512.

[Fellbaum 1998] Fellbaum, C. (1998; Eds.): *WordNet – An Electronic Lexical Database*. Cambridge & London: MIT Press.

[Mikk 2000] Mikk, J. (2000): *Textbook: Research and Writing*. Frankfurt a. M. u. a.: Lang.

[Saussure 1967] Saussure, F. de (^{2}1967): *Grundfragen der allgemeinen Sprachwissenschaft*. Berlin: Walter de Gruyter.

[Shannon 1948] Shannon, C. E. (1948): *A Mathematical Theory of Communication*. The Bell System Technical Journal, 27, pp. 379–423 & 623–656.