Extracting Informative Content Units in Text Documents[*]

Using Topic Chains for Conceptual Document Representation

Jürgen Reischer

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

Abstract
The notion of semantic and thematic informativeness of text is explored in theory and practice. The IVal system is described which implements a procedure for conceptual text analysis and representation based on thematic chains. Possible applications for semantic text processing including conceptual indexing and passage extraction are presented and discussed.

1 Introduction

The notion of *informativeness* has not gained much attention in the literature, although informativeness is the primary quality an information seeking person certainly expects of documents or passages retrieved. Contrary to the frequently discussed and defined concept of information,[1] the notion of informativeness lacks an explication which also impedes its application in computational language processing. For example, in order to detect and select informative content units in text documents, we need both a conception of informativeness and a procedure for automatic extraction of informative passages. In the following sections, therefore,

[*] Published in: OSSWALD, Achim; STEMPFHUBER, Maximilian; WOLFF, Christian (eds.) (2007). Open Innovation. Proc. 10. International Symposium for Information Science. Constance: UVK, 285-302.

[1] An extensive overview and discussion about ordinary and scientific concepts of information can be found in [Reischer 2006a].

we will address theoretical and practical aspects of informativeness in view of its explication and automatic extraction.

2 The Concept of Informativeness

We will approach the term 'informativeness' from both the perspective of everyday as well as scientific language. After that, we will give an explication of informativeness as used in this context and discuss related notions.

2.1 Informativeness in Everyday Language

The term 'informativeness', as derived from 'informative', may be understood from several perspectives if we consider its everyday meaning: on the one hand, an expression (sentence, text) may be informative *for itself* without reference to a certain interest or information need, e. g. the sentence "I am here now and doing something" is less informative than "J. R. works on his paper in Regensburg on 24/12/2006 at 15.00 o'clock"; on the other hand, a text (passage or document) may be informative *about* a certain topic or *relative to* an information need of a user (as expressed by search terms). Another distinction of 'informative(ness)' concerns the semiotic levels of syntax, semantics, and pragmatics. Primarily, informativeness is related to the *semantic* level of text in the sense of concepts or propositions denoted by signs. Syntactic informativeness may be understood, for example, as the information content of signs derived from their probability of appearance [cf. Shannon 1948], where 'sign' just means a symbolic *form* (irrespective of its meaning). Pragmatic informativeness, finally, can be interpreted as such instructive and/or enlightening semantic contents that immediately promote or prohibit action.

Besides these qualitative distinctions, informativeness is also a quantifiable concept, i.e. text may be *more* or *less* informative. On the one hand, this may simply mean that it contains more or less informational units measured absolutely or relatively to its length. On the other hand, it may be interpreted as the degree of specificity or preciseness of concepts and propositions conveyed by the text. In the first case, we have more pieces or bits of information; in the latter case, we simply gain more informational content.

2.2 Informativeness in Scientific Context

Departing from the ordinary understanding of 'informativeness', we will have a closer look at some uses of this notion in scientific contexts. Like many other terms, 'informativeness' is used ambiguously here; I will just give some examples. In Tague-Sutcliffe, informativeness simply means the *amount of information* conveyed by the documents or records provided by an information service [Tague-Sutcliffe 1995]. Qualitatively, information is to be understood here as conceptual content in accord with Fox' definition of information as propositions [cf. Fox 1983]. In the process of abstracting, informative summaries cover the *most important concepts*, as contained in corresponding passages of a text [cf. e. g. Mani & Maybury 1999]. In text linguistics, informativeness means the extent to which textual elements (words, sentences) are *expected* versus *unexpected* or *known* versus *unknown* [Beaugrande & Dressler 1981]. This explication seems like a semantic counterpart to Shannon's (syntactic) surprise value of signs. Rosch and colleagues consider informativeness of concepts from the perspective of basic level categories and their super- und subordinate concepts. In this regard, basic level concepts have optimal informativeness with respect to the *number of defining attributes* and the *number of attributes differentiating it from neighbour concepts*, thus minimizing the cognitive load for information processing [cf. Rosch & Mervis 1975; Rosch & al. 1976; Rosch 1978]. The author himself provides a measure of informativeness for concepts based on the assumption that the (vertical) position of a concept in a conceptual hierarchy (e. g. as realized in WordNet) is an indicator of its *semantic specificity* [Reischer 2006b]. The more specific a concept is, the lower it is positioned in the hierarchy and the more distinctive features it must carry, thus being more informative (as in 'This is a *collie*' versus 'This is a *dog*').

In summary, this short extract from the literature reveals a quite inhomogeneous usage of the term 'informativeness' which prevents a synopsis of the underlying conceptions. Nevertheless, the notion of informativeness primarily applies to semantic units like concepts as conveyed by signs (word, sentence, text), which may be more or less informative. The amount of information transmitted by linguistic expressions (on a certain topic) may be called the semantic-thematic or conceptual information content of a text. Because concepts are regarded as purely semantic units, they cannot be accessed immediately from the textual surface which consists merely of a sequence of *terms* in the sense of *morphosyntactical* units. Terms are assigned one or more concepts as their meaning (signification), so that these concepts can be accessed indirectly via the terms expressing them. Because many terms signify more than one concept, i.e. they are ambiguous, they have to be disambiguated to select the correct semantic information the term conveys in a certain (co)text.

In the next sections, we will give an explication of informativeness in the sense of semantic and thematic (= conceptual) information content.

2.3 An Explication of Informativeness

Informativeness could be understood as the potential of a text or a text passage of a document to satisfy a user's information need (better or worse). This may either be understood as the general or absolute capacity of a text to be informative as such (for any recipient in any context), or as the specific or relative capacity of a text to be informative about a certain topic or with respect to the information need of a user (as reflected by some search terms in either case). A text is more or less informative if it has the potential to convey a certain amount of semantic or thematic information (on a certain topic), i.e. it has a lower or higher information content. Necessarily the question raises, what properties of a text account for its informativeness.

Basically, semantic information is carried by single concepts or complex propositions constituted of concepts by forming predicate-argument structures. Although it would be highly desirable to compute propositional information contents automatically, we must restrict ourselves to the information contained in single concepts. With respect to computational aspects this can be realized with much less effort than completely parsing text for its syntax and semantics. Furthermore, conceptions of propositional information content in the literature are hardly translatable into adequate automatic procedures. In the following sections, therefore, we will go into details about the notions of computable semantic and thematic information content which can be used for practical applications.

2.3.1 Semantic Information Content

The notion of semantic information content may be tackled from two different perspectives: We may either look at properties of the concepts themselves or at the distribution of concepts within a text. In the first view, information content may be related to semantic specificity (as opposed to genericity) or semantic frequency of occurrence (of concepts); in the latter view, semantic information can be related to semantic density.

The notion of semantic specificity versus genericity of concepts is based on the idea that a concept contains the more information the less other concepts it classifies (and vice versa). For example, the concept ENTITY classifies all other concepts, because everything is an entity; in contrast, the concepts SEPTEMBER-11-2001 or ALBERT-EINSTEIN classify no further concepts, because they are informationally as discriminative as possible denoting only one specific thing in the world. On the linguistic level, a hyponym (subordinate) of a term carries more information than its hyperonym (superordinate). For instance, 'collie' – or better: its meaning – is more specific than 'dog', and 'Lassie' is more specific than 'collie'. If I am said that X is a collie, I gain more knowledge than if I hear that X is an animal. In order to automatically compute information contents for concepts in the sense of semantic specificity, we need a conceptual hierarchy with hyperonym-hyponym relations between these concepts. One such hierarchy is available with WordNet (or GermaNet or any other net), where nouns and verbs are organised into hierarchical trees from the most generic to the most specific concepts. An algorithm for computing information values on the basis of the vertical position of a concept in the hierarchy can be found in [Reischer 2006b].

In an alternative measure of semantic information, the frequency account of Shannon's information content (Shannon 1948) is combined with the above approach based on conceptual hierarchies. The basic assumption behind this approach is twofold: firstly, the observation that frequently recurring entities carry less information than rarely occurring entities (compare the more or less probable symbols of an alphabet in Shannon's theory);[2] secondly, the occurrence of a semantically more specific concept X necessarily entails the *conceptual* cooccurrence of a more general concept Y superordinated to X.[3] That is to say, if we talk of collies, we also talk of dogs; if we think of Lassie, we also think of collies. Consequently, all occurrences of concepts X_Y subordinated to Y contribute to the *semantic* occurrence of Y, because we cannot think of X_Y without thinking Y. The semantic information content of Y, then, is in analogy to Shannon, the binary logarithm of the frequency F(Y) of a concept in question *including the frequencies of all its subordinate concepts X_Y* in relation to the frequency of the upmost root concept F(R) including the summed-up frequencies of all its subordinates X_R: $I(Y) = -\log_2 (F(Y) / F(R))$, where both numerator and denominator must be greater than 0. For normalization, we must divide by the maximum information value possible (i.e. if a concept has frequency 1): $I_N(Y) = [\log_2 (F(Y) / F(R))] / [\log_2 (1 / F(R))]$. For $F(Y) = 1$, we get $I_N(Y) = 1$ (maximum information content), for $F(Y) = F(R)$, we get 0 (minimum information content). A low information content means that a concept Y occurs frequently, both by direct activation via an appropriate term (e. g. 'thing' etc.) and by being activated indirectly via all subordinated concepts and their respective terms (e. g. 'dog', 'collie', 'Lassie' etc.). The higher the frequency of a concept Y the

[2] In this approach, we are dealing with frequencies of *concepts* as semantic units, not with terms as symbols in the sense of morphosyntactic units.

[3] See Resnik (1995) for a comparable approach in semantic similarity measurement. The *concept* frequencies required can be derived from the WordNet database where the frequencies of terms *in certain senses* (i.e. concepts) are available.

more subordinates X_Y it typically has contributing to the frequency of Y, i.e. the more upward in the conceptual hierarchy and the more unspecific it must be. Necessarily, such a concept occurs in many communicative contexts and, therefore, must be undiscriminative with respect to the information it carries (because it has to be compatible with many other concepts in order to be semantically combinable with them).[4]

The notion of semantic density or concentration may be explicated simply by the ratio of the number N_C of content words to the number of function words N_F in a text or passage with $N_C + N_F$ words: $D = N_C / (N_F + N_C)$. The more content elements a text contains, the less uninformative (formative) units it can have (at equal length). The ratio is 1 if all items in the passage are content words, as in telegrams like 'Coming Sunday morning John'; the ratio is 0 in sentences with only functions words like 'I am here now'. Content words are typically nouns, verbs, adjectives, and many adverbs; function words are basically prepositions, conjunctions, pronouns, and determiners.[5] This quite simple measure of semantic informativeness is based on the idea that a text (passage) cannot convey more semantic information than is expressed by content words.

2.3.2 Thematic Information Content

Thematic informativeness is basically concerned with the topic structure of a text. A text is thematically more or less informative, if it conveys one or more topics in a more or less cohesive/coherent and concentrated (dense) way. Cohesion and coherence concern the syntactic and semantic connectedness of textual units (terms and concepts), respectively, to form a thematic whole [cf. Beaugrande & Dressler 1981]. Concentration means the density of these units with respect to a certain length of a textual passage.

The notions of cohesion and coherence are related to the syntactic and semantic means a text (writer) uses to form an interconnected sequence of words and sentences, concepts and propositions [cf. e. g. Halliday & Hasan 1976; Hoey 1991; Morris & Hirst 1991]. Thematic convergence of these textual units guarantees that the text is not an accidental, senseless concatenation of words and sentences but forms a consistent whole containing and conveying one or more topics. On the co-

[4] The content of this paragraph is taken from [Reischer 2007b].

[5] It is certainly debatable whether all prepositions are really function words, e. g. those derived from verbs like 'corresponding to' and 'according to'. Another point is the question in how far pronouns are not informative: if they are treated like the nouns they stand for, they *are* of course informative.

hesion level, this is accomplished by at least two syntactic means: pronouns refer back to thematic units already introduced (functioning like definite noun phrases); collocations (in the two senses of lexicalised multi-word expressions like idioms and otherwise contextually co-occurring terms) connect isolated words to meaningful clusters. On the coherence level, two phenomena connect concepts and propositions: semantic coreference and thematic associations between single concepts constitute chains of conceptually related elements (e. g. car – automobile, car – steering-wheel), conjunctions between propositions relate one complex content to another. The more cohesive and coherent units a text contains, the more thematically convergent it is assumed to be due to its formal and conceptual interwovenness. Computationally, this may be realized by a procedure assembling conceptually related elements to thematic chains: the length and strength of such chains are an indicator for the thematic information content (topicality) of the text (see section 3.1 below for further elaboration).

Thematic concentration or density means that some conceptual units of a text are such concepts constituting a topic of the text sustaining many thematic (conceptual) relations to other units in the same passage or text. Furthermore, the average thematic distance between all concepts is comparatively low so that many or even all conceptual units are strongly interconnected to a thematic network leaving only few concepts isolated. For example, the relatedness between concepts like INTERNET, FIREFOX, and BROWSER is very strong, whereas the connection between the concepts INTERNET, CELLPHONE, and CABLE is somewhat looser; in contrast, the concepts INTERNET, LASSIE, and JUSTICE constitute no intrinsic thematic relation. Another indicator of thematic concentration may be the type-token distribution of the text's concepts: if a text has few concepts with high frequency and many concepts with low frequency (relative to the total number of concepts), it is thematically focussed; if we have a text with a more homogenous type-token profile, the topic seems more diffuse or distributed. However, in both cases we have to consider 'thematic noise', i.e. the relative number of thematically singular and unconnected concepts being isolated from all others (see below).

2.4 Some Annotations to the Notion of Informativeness

Informativeness, understood as semantic or thematic information content, is intrinsically related to the notions of text comprehensibility and aboutness or relevance. Some details may be indicated in the following.

With respect to semantic understandability, as opposed to the linguistic surface phenomena of legibility and readability, a text is the less comprehensible the more (new) information it codes in a certain passage of text, be it the concentration of content units (propositional density) per formal unit, conceptual specificity (optimal basic concept level), or the explicit presence of sentence connectives [cf. e. g. Rosch & al. 1976; Kintsch & Vipond 1979]. That is to say, if a text is hard to understand, the probability of misunderstanding or even non-understanding increases so that the text will not be as informative for the recipient as potentially possible. Under this view, we have to discriminate between objective and subjective informativeness: the amount of information coded in the text is its theoretical *potential* that can be conveyed, the amount of information decoded by the recipient is the *actual* information arriving at the user. Consequently, if we want to quantify subjective information content, we have to consider general cognitive factors of comprehensibility (e. g. frequent and/or basic level concepts are even understandable by children) as well as specific factors of the recipient (e. g. his preknowledge of a domain). We are primarily concerned here with objective informativeness because subjective informativeness necessarily depends on the latter; i.e. the recipient cannot extract more semantopragmatic information than is semantically coded in the text.

The concept of relevance and its different interpretations in information science was summarized and discussed in [Mizzaro 1997]. As a common ground, relevance is a relational concept: something is relevant *for* something else. In contrast, informativeness can be absolute or relative: something is informative (as such) versus informative for/about something else. In the latter case, informativeness-for/about includes relevancy: a document is informative about a certain topic/question or for the information need of a user, if it is at least (thematically) relevant to the latter *and* if it is understandable and qualitatively well-built (e. g. orthographically well-formed, thematically coherent). Thus, a text document rated as relevant is actually not relevant for the user if it is completely incomprehensible or obviously lacks quality control. The user wants informative documents, not just relevant ones. This should be considered at least in passage or document ranking.

3 Realization of Informativeness Evaluation

The automatic detection and selection of informative contents in text documents can serve two different goals: firstly, in the text-oriented view, we want to extract those contents which are most 'representative' for the content of a text, i.e. an index of concepts (and respective terms representing them) or a list of passages being most informative about the topic(s) of the entire document (indexing and summarizing); secondly, in the user-oriented view, we want to extract those sections of the text that match best with the required contents expressed by one or more search terms representing the concepts desired (sentence/section retrieval). In addition to the detection and selection of the most informative passages with respect to thematic content, we may further select those passages from the list of candidates with the highest/lowest semantic information contents (ranking of passages, e. g. for comprehensibility, conciseness, detailedness, etc.).

The two scenarios presented above unify several information processing tasks to one general conception of generally detecting and selecting informative contents in text documents: summarizing is viewed as the extraction of the most informative passages based on one or more concepts being thematically most representative for the entire text (retrieval by the most important concepts of the text taken from its concept index); passage retrieval is conceptionally based on the same mechanism replacing internal by external, user-defined search terms and their associated concepts. Thus, we can efficiently implement summarizing (as extraction of sections) and passage retrieval with the same procedure to be described in the next section.

3.1 Implementation

In the following subsections, the first implementation of a general procedure for semantic informativeness extraction on the basis of the WordNet and IVal systems and their applications will be described in more detail.

3.1.1 The WordNet Lexical Database

The automatic detection of informative contents in text documents using concepts instead of terms presupposes the existence of machine-readable lexical resources both containing terms and their associated concepts. One such resource is WordNet, which is a freely available lexical database in an easy-to-use text format [Fellbaum 1998]. In version 2.1, it contains about of 150.000 content terms (including about 64.000 multi-word terms), 120.000 concepts, 200.000 associations between terms and concepts (as senses of the terms), and 300.000 conceptual relations between concepts. The semantic network comprises relations like synonymy, antonymy, hyperonymy, hyponymy, meronymy, holonymy, and some others. WordNet has been widely used for linguistic processing of texts, e. g. for summarizing and lexical chaining [Barzilay & Elhadad 1997] as well as indexing [Gonzalo & al. 1998].

3.1.2 The IVal System

IVal[6] is an experimental system designed to provide access to the WordNet lexical database and to use the vast amount of linguistic and conceptual knowledge coded there for text analysis [Reischer 2007a]. The system in its current state consists of several text processing components:

- sentence boundary detection for the dissection of input text into single sentences;
- morphological parser for deflection and decomposition of simplex and complex terms (derivatives and compounds);
- chart parser based multi-word term recognizer for assembling collocations and idioms;
- simple proper name analyser for uninterpretable expressions not recognized by other analysis modules;
- interactive interface to the WordNet database for the expansion of lexical terms and concepts as well as relations between them (domain modeller).

Additionally, the WordNet lexicon has been enlarged for function words and term frequency data. The basic architecture of the system is task-oriented: the browser provides extended access to the WordNet lexical database; the weaver enables the user to define new terms, concepts, and associations; the analyser reads and analyses plain texts.

3.1.3 Thematic Chains

Topic or thematic chains are an extension of lexical chains [e. g. Barzilay & Elhadad 1997; Silber & McCoy 2002]. Lexical chaining uses the conceptual interwovenness of lexical entries to find semantic relations between terms (or better their associated concepts) in a text. The result is a chain of coreferent terms which represent those concepts of a text that are not only repeated but semantically linked. The strongest chains with respect to length (number of coreferents) and strength (semantic distance/similarity/proximity between the concepts) are a good indication of what a text is about on the conceptual level. In most cases, only synonym, hyperonym, and hyponym relations are used to form coreference chains. However, as [Varelas & al. 2005] and [Budanitsky & Hirst 2006] pointed out, further conceptual relations like antonymy, meronymy, holonymy, and others should be considered. In that case, we must talk of *thematic or topic chains* because the related concepts are not just more or less semantically equal but thematically related. One example may show this: The

[6] *I*nformativeness *Eval*uator.

shortest *semantic* distance between the concepts AUTOMOBILE and STEERING-WHEEL is 10, if we simply count all intervening nodes in WordNet 2.1 to step from one concept to the other; in contrast, if we use more relations like the ones given above, the distance reduces to 3.[7]

The automatic construction of thematic chains is basically quite simple: For every identified noun term in the text, all its possible readings (i.e. associated concepts) are tried to be thematically linked to other concepts in existing and open chains. A concept is linked to a chain if its thematic distance to the last N concepts of the chain is below a certain threshold. Concepts not linkable to any existing chain open new chains; chains to which no concept could be linked for a certain period are closed. A chain grows and gets stronger the more thematic relations it can establish to other text concepts; a chain dies if it contains only concepts which are not further supported by the cotext (then it was a wrong thematic thread due to an inadequate interpretation of a term). For example, if the new concept COMPUTER had to be chained to two possible existing chains <PC, NOTEBOOK> or <KEYBOARD, LCD, DRIVE> then there is a strong tendency to link it to one of the chains. But the possible sense 'an expert at calculation' of 'computer' is certainly not linked to any of these chains because it is simply the wrong reading in that context.

In more detail, the procedure for thematic chaining includes several steps:

- Selection of appropriate terms: (i) possible nouns are identified from the text; if a term is ambiguous between two parts of speech (e. g. 'American'), then the term is used as noun because the concepts are strongly related; (ii) very polysemous and frequent noun terms are excluded from further consideration because they seem semantically not discriminative enough (e. g. 'thing'), i.e. they are somehow related to any concept; (iii) for the same reason, concepts of a term having low semantic information contents are excluded [see section 2.3.1]: they are too unspecific or vague (e. g. 'man' in the sense of ADULT MALE is excluded but not as GAME EQUIPMENT CONSISTING OF AN OBJECT USED IN CERTAIN BOARD GAMES).
- Comparison of concepts to chains: the noun concept to be chained is compared against every existing chain and the concepts already included there. For that purpose, the average thematic proximity between the new concept and the concepts already chained has to be measured. If the distance is below a certain

[7] The reader may inspect the conceptual path between the two concepts by entering "? automobile & steering wheel" for semantic and "? automobile @ steering wheel" for thematic distance in the IVal browser.

threshold, then the concept is linked to the chain, i.e. extends it. Thematic proximity is calculated as the extent of weighted feature overlap between two concepts, where the features are either conceptually linked neighbour concepts describing the semantothematic vicinity of the two concepts in question, respectively, or the normalized terms extracted from their synsets and glosses.

- Scoring and sorting of chains: the finally resulting chains have to be scored by their length and strength, i.e. the average proximity between the concepts in the chain and the number of concepts included there. A simple measure is the product of these two values (if proximity is maximal at unity). Certain chains must be excluded from the list: e. g. chains containing only one element obviously sustain no relations to other concepts of the text and can be considered as thematic noise. After scoring and sorting out, the most important or representtative concepts of the text can be extracted by simply counting the number of their occurrences in all chains. The justification for this procedure is the fact that only concepts with many thematic relations to other concepts are frequently chained so that they must be central to the topic of the text.

The performance of thematic chaining, both quantitatively with respect to processing time and qualitatively with respect to the results obtained, depends heavily on several parameters used in the chaining process. These parameters include

- the maximum frequency and degree of polysemy a noun term may have (relative to the most frequent and polysemous noun) to consider its concepts in chaining at all;
- the minimum semantic information content a concept must have to be used for chaining at all;
- the maximum distance for thematic proximity of two concepts as well as the scoring of the different conceptual relation types (e. g. a hyperonym concept of X is thematically closer to X than a meronym concept of X);
- the maximum number of steps no concept is linked to a chain until it is closed again.

The optimal parameter setting can only be found by experimentation and evaluation which is work in progress; nevertheless, reasonable initial values can be set intuitively to be later adjusted. However, one result is obvious from the beginning: the qualitative performance depends on the density of the conceptual network modelled. The more conceptual relations are available in the lexicon for a certain domain, the better thematic analysis and assessment of texts can be.

Consequently, for further performance improvement additional knowledge resources have to be integrated (see section 3.2 below).

Using thematic chaining for conceptual representation of texts has several advantages:

- Ambiguous terms can be disambiguated: For example, the term 'apple' has several meanings in English, but in the conceptual neighbourhood of MICROSOFT (within one and the same conceptual chain) it is probably to be interpreted as APPLE COMPANY, because thematic proximity between MICROSOFT and APPLE COMPANY is certainly higher than between MICROSOFT and APPLE FRUIT/TREE.
- Automatic segmentation of text: Every thematic chain represents a certain span of text, where more than one chain may cover one and the same passage (due to different interwoven topics within a section). We can simply count the number of chains starting or ending at a certain text position (e. g. at a certain sentence), so that we have a good indication where (sub)topics start and end.
- Rating of concept importance: Concepts appearing in more than one chain sustain several thematic relations to other concepts and may be deemed as central concepts. The more often a concept appears in a chain the more important and central it is. Indirectly, the frequency of a concept is considered because this increases the possibility to be thematically linked to other concepts.
- Rating of text coherence: The less chains a text needs for representation, the more thematically coherent and dense it seems to be. This may be regarded as a quality criterion for the text, e. g. a text is the more understandable the more coherent it is.

In the next section, we will look at some applications of thematic chaining as one form of text representation for informativeness extraction.

3.1.4 Applications

Detecting and selecting informative contents from a text affects at least three scenarios: extraction of the most important concepts with respect to the central topic(s) of the text (conceptual indexing), extraction of the most informative or most representative passages of the text (informative summary on the basis of the most important concepts), and retrieval of the most informative sections relative to a set of search terms (passage retrieval; see e. g. [Salton & al. 1993]).

Conceptual indexing as extraction of the most important concepts of the text is required primarily for summarizing here: If we take the N most significant concepts and perform a passage retrieval operation (see below) then we gain the most informative sections (as defined by the text span of a thematic chain) containing the most important concepts. Furthermore, we can use the concept index to reversely create a term index: in WordNet, every concept is linked to a so-called synset (synonym set) which contains all possible terms expressing the concept. The advantage of this procedure is the inclusion of real topic terms (based on topic concepts) automatically excluding thematic noise terms.

The retrieval of passages being informative about some search terms and their concepts employs the same algorithm as used in thematic chaining: all search terms in their possible readings (concepts) are tried to be included in every thematic chain by conceptual distance. The more of the concepts of the terms can be (virtually) chained with the less thematic distance, the better a specific interpretation (combination of concepts) of the search terms matches with a chain. For example, the search terms 'keyboard' and 'drive' have both several possible meanings of which only the combination COMPUTER-KEYBOARD and PC-STORAGE-DRIVE are thematically or conceptually close to the chain <KEYBOARD, LCD, DRIVE>; neither other term readings (e. g. KEY-HOLDER and DRIVEWAY) nor other chains (e. g. <PC, NOTEBOOK>) are appropriate candidates for adequate results.

In ranking, the passages are primarily scored by their total average distance of all terms' best readings. Additionally, the most informative passages with respect to semantic and thematic information content may be scored for understandability (e. g. semantic specificity or density) and local concentration (how many of the text's most important topic terms are included in the passage). The performance of informativeness detection and selection as described above is still a matter of evaluation. First results indicate an approach worth to be further explored. Nevertheless, further experiments have to be conducted: they concern the question, which parameters of the chaining and retrieval process yield optimal performance, because calibration and coordination of all parameters is of major importance for performance. Another question to be answered is the gain in performance by different improvements like pronoun resolution, proper name recognition, better thematic distance measures, as well as WordNet expansion of terms, concepts, and thematic relations.

3.2 Prospects

Of course, there are many questions open. We are just at the beginning of automatic detection and selection of informative contents. The transition from formal to content-related text processing is unavoidable: with the advent of the semantic web we need new and better strategies to exploit the conceptual-thematic information coded in wordnets and ontologies. Terminologically and conceptionally, we should replace the notion 'relevant' by 'informative' indicating the semantic approach to text processing. With ontologies and wordnets becoming continuously larger and better, the retrieval performance will necessarily increase. At the time being, lexical and ontological resources are still far from being perfect with respect to the density of the general and specific knowledge about topic domains. This also limits the application of the procedures described above to well-modelled subdomains. However, in the near future we will have a complete and up-to-date coverage of all terms and concepts in the world, structured into an ontology or WordNet. If such a complete ontology is available it will not vanish anymore.

One possibility to compensate for the current shortcomings is the automatic enlargement of wordnets and ontologies with respect to additional thematic relations. For example, other knowledge resources like the Cyc ontology may be merged with WordNet and enhance qualitative performance. Quantitative performance can be boosted by a complete precalculation of all thematic relation combinations of noun concepts in WordNet (about 80.000^2 / 2 concept-concept relations, with 1 byte per pair need 3 GB of memory). The future generations of multicore CPUs and large memories will increase performance drastically.[8] As a consequence, informativeness evaluation of whole documents and document collections will be possible on every PC.

4 Conclusion

The extraction of informative content units in documents by semantic text analysis is still in its infancy. The IVal system implements one possible approach to concept-based text representation and passage retrieval by exploiting thematic relations between concepts. One benefit of this approach is the effectiveness of implementation: if the procedure for thematic chaining is implemented, we automatically gain a conceptual indexer, summarizer, and passage retriever which are all based on thematic distance measuring. Further experiments must show whether this approach can also be scaled up to document collections.

[8] The performance of syntactic term based search cannot be increased anymore, as Google's immediate presentation of retrieval results in the largest document collection of the world impressively proves. The additional processing power should be better invested in semantic search engines.

5 Bibliography

[Barzilay & Elhadad 1997] Barzilay, R. & Elhadad, M. (1997): Using lexical chains for text summarization. In *Proceedings of the Intelligent Scalable Text Summarization Workshop (ISTS'97)*, ACL. *http://citeseer.ist.psu.edu/article/barzilay97using.html* (8.1.2007)

[Beaugrande & Dressler 1981] Beaugrande de, R.-A. & Dressler, W. (1981): *Introduction to Text Linguistics*. London & New York: Longman.

[Budanitsky & Hirst 2006] Budanitsky, A. & Hirst, G. (2006): Evaluating WordNet-based Measures of Lexical Semantic Relatedness. *Computational Linguistics*, 32(1). pp. 13 – 47.

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

[Gonzalo & al. 1998] Gonzalo, J. & Verdejo, F. & Chugur, I. & Cigarran, J. (1998): Indexing with WordNet synsets can improve Text Retrieval. *Proceedings of the COLING/ ACL '98 Workshop on Usage of WordNet for NLP*. *http://citeseer.ist.psu.edu/article/gonzalo98indexing.html* (8.1.2007)

[Fox 1983] Fox, C. J. (1983): Information and Misinformation. An Investigation of the Notions of Information, Misinformation, Informing, and Misinforming. Westport & London: Greenwood Press.

[Halliday & Hasan 1976] Halliday, M. A. K. & Hasan, R. (1976): *Cohesion in English*. London & New York: Longman.

[Hoey 1991] Hoey, M. (1991): *Patterns of Lexis in Text*. Oxford: University Press.

[Kintsch & Vipond 1979] Kintsch, W. & Vipond, D. (1979): Reading Comprehension and Readability in Educational Practice and Psychological Theory. In Nilsson, L.-G. (Eds.): *Perspectives on Memory Research*. Hillsdale: Erlbaum. pp. 329 – 365.

[Mani & Maybury 1999] Mani, I. & Maybury, M. T. (1999; Eds.): *Advances in Automatic Text Summarization*. Cambridge & London: MIT Press.

[Mizzaro 1997] Mizzaro, S. (1997): Relevance: The Whole History. *JASIS*, 48(9). pp. 810 – 832.

[Morris & Hirst 1991] Morris, J. & Hirst, G. (1991): Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of Text. *Computational Linguistics*, 17(1). pp. 21 – 48.

[Reischer 2006a] Reischer, J. (2006): *Zeichen Information Kommunikation. Analyse und Synthese des Zeichen- und Informationsbegriffs*. *http://www.opus-bayern.de/uni-regensburg/volltexte/2006/740/pdf/ZeichenInfoKomm.pdf* (accessed 8.1.2007)

[Reischer 2006b] Reischer, J. (2006): IVal – An Alternative WordNet Browser for Evaluating Semantic Informativeness of Concepts. Proceedings der KONVENS 2006, pp. 115 – 120. *http://ling.uni-konstanz.de/pages/conferences/konvens06/konvens_files/konvens06-proc.pdf* (accessed 14.10.2006)

[Reischer 2007a] Reischer, J. (2007): *IVal – Informativeness Evaluator for Retrieval*. *http://lingua-ex-machina.de* (accessed 8.1.2007)

[Reischer 2007b] Reischer, J. (2007): OntoNet – a WordNet-based ontological-lexical development system. (To appear in the Proceedings of the GLDV-07 Workshop on Lexical-Semantic and Ontological Resources, 13.4. – 14.4.2007, Tübingen)

[Resnik 1995] Resnik, P. (1995): Using Information Content to Evaluate Semantik Similarity in a Taxonomy. *Proceedings of the IJCAI-95*, Vol. I. pp. 448 – 453.

[Rosch & Mervis 1975] Rosch, E. & Mervis, C. B. (1975): Family Resemblances: Studies in the Internal Structure of Categories. *Cognitive Psychology*, 7. pp. 573 – 605.

[Rosch & al. 1976] Rosch, E. & Mervis, C. B. & Gray, W. D. & Johnson, D. M. & Boyes-Braem, P. (1976): Basic Objects in Natural Categories. *Cognitive Psychology*, 8. pp. 382 – 439.

[Rosch 1978] Rosch, E. (1978): Principles of Categorization. In Rosch, E. & Lloyd, B. B. (1978; Eds.): *Cognition and Categorization*. Hillsdale: Erlbaum. pp. 27 – 48.

[Salton & al. 1993] Salton, G. & Allan, J. & Buckley, C. (1993): Approaches to passage retrieval in full text information systems. In *ACM SIGIR conference on R&D in Information Retrieval*. pp. 49 – 58. *http://citeseer.ist.psu.edu/salton93approaches.html* (8.1.2007)

[Shannon 1948] Shannon, C. E. (1948): A Mathematical Theory of Communication. The Bell System Technical Journal, 27. pp. 379 – 423 & 623 – 656. *http://cm.belllabs.com/cm/ms/what/shannonday/shannon1948.pdf* (accessed 14.10.2006)

[Silber & McCoy 2002] Silber, H. G. & McCoy, K. F. (2002): Efficiently Computed Lexical Chains as an Intermediate Representation for Automatic Text Summarization. *Computational Linguistics*, 28(4). pp. 487 – 496.

[Tague-Sutcliffe 1995] Tague-Sutcliffe, J. (1995): *Measuring Information. An Information Services Perspective*. San Diego u a.: Academic Press.

[Varelas & al. 2005] Varelas, G. & Voutsakis, E. & Raftopoulou, P. (2005): Semantic Similarity Methods in WordNet and their Application to Information Retrieval on the Web. WIDM'05. pp. 10 – 16.