Abstraction Summarization for Managing the Biomedical Research Literature

Marcelo Fiszman **Thomas C. Rindflesch** **Halil Kilicoglu**

Lister Hill National Center for Biomedical Communications
National Library of Medicine
Bethesda, MD 20894
{fiszman|tcr|halil}@nlm.nih.gov

Abstract

We explore a semantic abstraction approach to automatic summarization in the biomedical domain. The approach relies on a semantic processor that functions as the source interpreter and produces a list of predications. A transformation stage then generalizes and condenses this list, ultimately generating a conceptual condensate for a disorder input topic. The final condensate is displayed in graphical form. We provide a set of principles for the transformation stage and describe the application of this approach to multidocument input. Finally, we examine the characteristics and quality of the condensates produced.

1 Introduction

Several approaches to text-based information management applications are being pursued, including word-based statistical processing and those depending on string matching, syntax, or semantics. Statistical systems have enjoyed considerable success for information retrieval, especially using the vector space model (Salton et al., 1975). Since the SIR system (Raphael, 1968), some have felt that automatic information management could best be addressed using semantic information. Subsequent research (Schank, 1975; Wilks, 1976) expanded this paradigm. More recently, a number of examples of knowledge-based applications show considerable promise. These include systems for machine translation (Viegas et al., 1998), question answering, (Harabagiu et al., 2001; Clark et al., 2003), and information retrieval (Mihalcea and Moldovan, 2000).

In the biomedical domain, the MEDLINE® bibliographic database provides opportunities for keeping abreast of the research literature. However, the large size of this online resource presents potential challenges to the user. Query results often include hundreds or thousands of citations (including title and abstract). Automatic summarization offers potential help in managing such results; however, the most popular approach, extraction, faces challenges when applied to multi-document summarization (McKeown et al., 2001).

Abstraction summarization offers an attractive alternative for managing citations resulting from MEDLINE searches. We present a knowledge-rich abstraction approach that depends on underspecified semantic interpretation of biomedical text. As an example, a graphical representation (Batagelj, 2003) of the semantic predications serving as a summary (or conceptual condensate) from our system is shown in Figure 1. The input text was a MEDLINE citation with title "Gastrointestinal tolerability and effectiveness of rofecoxib versus naproxen in the treatment of osteoarthritis: a randomized, controlled trial."

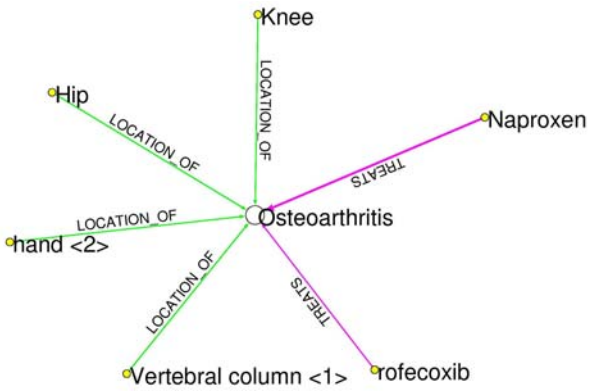

Figure 1. Semantic abstraction summarization

Our semantic interpreter and the abstraction summarizer based on it both draw on semantic information from the Unified Medical Language System® (UMLS),® a resource for structured knowledge in the biomedical domain. After introducing the semantic interpreter, we describe the transformation phase of our paradigm, discussing principles that depend on semantic notions in order to condense the semantic predications representing the content of text. Initially, this process was applied to summarizing single documents. We discuss its adaptation to multidocument input, specifically to the set of citations resulting from a query to the MEDLINE database. Although we have not yet formally evaluated the effectiveness of the resulting condensate, we discuss its characteristics and possibilities as both an indicative and informative summary.

2 Background

2.1 Lexical Semantics

Research in lexical semantics (Cruse, 1986) provides insight into the interaction of reference and linguistic structure. In addition to paradigmatic lexical phenomena such as synonymy, hypernymy, and meronymy, diathesis alternation (Levin and Rappaport Hovav, 1996), deep case (Fillmore, 1968), and the interaction of predicational structure and events (Tenny and Pustejovsky, 2000) are being investigated. Some of the consequences of research in lexical semantics, with particular attention to natural language processing, are discussed by Pustejovsky et al. (1993) and Nirenburg and Raskin (1996). Implemented systems often draw on the information contained in WordNet (Fellbaum, 1998).

In the biomedical domain, UMLS knowledge provides considerable support for text-based systems. (Burgun and Bodenreider (2001) compare the UMLS to WordNet.) The UMLS (Humphreys et al., 1998) consists of three components: the Metathesaurus,® Semantic Network (McCray, 1993), and SPECIALIST Lexicon (McCray et al., 1994). The Metathesaurus is at the core and contains more than 900,000 concepts compiled from more than sixty controlled vocabularies. Many of these have hierarchical structure, and some contain meronymic information in addition to hypernymy. Editors combine terms in the constituent vocabularies into a set of synonyms (cf. WordNet's synsets), which constitutes a concept. One term in this set is called the "preferred name" and is used as the concept name, as shown in (1).

(1) **Concept**: Dyspnea **Synonyms**: Breathlessness, Shortness of breath, Breathless, Difficulty breathing, Respiration difficulty, etc.

In addition, each concept in the Metathesaurus is assigned at least one semantic type (such as **'**Sign or Symptom' for (1)), which categorizes the concept in the biomedical domain. The semantic types available are drawn from the Semantic Network, in which they are organized hierarchically in two single-inheritance trees, one under the root 'Entity' and another under 'Event'.

The Semantic Network also contains semantic predications with semantic types as arguments. The predicates are semantic relations relevant to the biomedical domain and are organized as subtypes of five classes, such as TEMPORALLY_RELATED_TO and FUNCTIONALLY_RELATED_TO. Examples are shown in (2).

(2) **'**Pharmacologic Substance' TREATS 'Disease or Syndrome', 'Virus' CAUSES 'Disease or Syndrome'

Lexical semantic information in the UMLS is distributed between the Metathesaurus and the Semantic Network. The Semantic Network stipulates permissible argument categories for classes of semantic predications, although it does not refer to deep case relations. The Metathesaurus encodes synonymy, hypernymy, and meronymy (especially for human anatomy). Synonymy is represented by including synonymous terms under a single concept. Word sense ambiguity is represented to some extent in the Metathesaurus. For example *discharge* is represented by the two concepts in (3), with different semantic types.

(3) **Discharge, Body Substance**: 'Body Substance'
Patient Discharge: 'Health Care Activity'

The SPECIALIST Lexicon contains orthographic information (such as spelling variants) and syntactic information, including inflections for nouns and verbs and sub-categorization for verbs. A suite of lexical access tools accommodate other phenomena, including derivational variation.

2.2 SemRep

Our summarization system relies on semantic predications provided by SemRep (Rindflesch and Fiszman, 2003), a program that draws on UMLS information to provide underspecified semantic interpretation in the biomedical domain (Srinivasan and Rindflesch, 2002; Rindflesch et al., 2000). Semantic interpretation is based on a categorical analysis that is underspecified in that it is a partial parse (cf. McDonald, 1992). This analysis depends on the SPECIALIST Lexicon and the Xerox part-of-speech tagger (Cutting et al., 1992) and provides simple noun phrases that are mapped to concepts in the UMLS Metathesaurus using MetaMap (Aronson, 2001).

The categorial analysis enhanced with Metathesaurus concepts and associated semantic types provides the basis for semantic interpretation, which relies on two components: a set of "indicator" rules and an (underspecified) dependency grammar. Indicator rules map between syntactic phenomena (such as verbs, nominalizations, and prepositions) and predicates in the Semantic Network. For example, such rules stipulate that the preposition *for* indicates the semantic predicate TREATS in *sumatriptan for migraine.* The application of an indicator rule satisfies the first of several necessary conditions for the interpretation of a semantic predication.

Argument identification is controlled by a partial dependency grammar. As is common in such grammars, a general principle disallows intercalated dependencies (crossing lines). Further, a noun phrase may not be used as an argument in the interpretation of more than one semantic predication, without license. (Coordination and relativization license noun phrase reuse.) A final principle states that if a rule can apply it must apply.

Semantic interpretation in SemRep is not based on the "real" syntactic structure of the sentence; however linear order of the components of the partial parse is crucial. Argument identification rules are articulated for each indicator in terms of surface subject and object. For example, subjects of verbs are to the left and objects are to the right. (Passivization is accommodated before final interpretation.) There are also rules for prepositions and several rules for arguments of nominalizations.

The final condition on the interpretation of an associative semantic predication is that it must conform to the appropriate relationship in the Semantic Network. For example, if a predication is being constructed on the basis of an indicator rule for TREATS, the syntactic arguments identified by the dependency grammar must have been mapped to Metathesaurus concepts with semantic types that conform to the semantic arguments of TREATS in the Semantic Network, such as 'Pharmacologic Substance' and 'Disease or Syndrome'. Hypernymic propositions are further controlled by hierarchical information in the Metathesaurus (Rindflesch and Fiszman, 2003).

In processing the sentence in (4), SemRep first constructs the partial categorical representation given schematically in (5). This is enhanced with semantic information from the Metathesaurus as shown in (6), where the corresponding concept for each relevant noun phrase is shown, along with its semantic type. The final semantic interpretation for (4) is given in (7).

(4) Mycoplasma pneumonia is an infection of the lung caused by Mycoplasma pneumoniae

(5) [[Mycoplasma pneumonia] [is] [an infection] [of the lung] [caused] [by Mycoplasma pneumoniae]]

(6) "Mycoplasma pneumonia"–'Disease or Syndrome'
"Infection"–'Disease or Syndrome'
"Lung"–'Body Part, Organ, or Organ Component'
"Mycoplasma pneumoniae"–'Bacterium'

(7) Mycoplasma Pneumonia ISA Infection
Lung LOCATION_OF Infection
Lung LOCATION_OF Mycoplasma Pneumonia
Mycoplasma pneumoniae CAUSES Infection
Mycoplasma pneumoniae CAUSES Mycoplasma Pneumonia

3 Automatic Summarization

Automatic summarization is "a reductive transformation of source text to summary text through content reduction, selection, and/or generalization on what is important in the source" (Sparck Jones, 1999). Two paradigms are being pursued: extraction and abstraction (Hahn and Mani, 2000). Extraction concentrates on creating a summary from the actual text occurring in the source document, relying on notions such as frequency of occurrence and cue phrases to identify important information.

Abstraction, on the other hand, relies either on linguistic processing followed by structural compaction (Mani et al., 1999) or on interpretation of the source text into a semantic representation, which is then condensed to retain only the most important information asserted in the source. The semantic abstraction paradigm is attractive due to its ability to manipulate information that may not have been explicitly articulated in the source document. However, due to the challenges in providing semantic representation, semantic abstraction has not been widely pursued, although the TOPIC system (Hahn and Reimer, 1999) is a notable exception.

3.1 Semantic Abstraction Summarization

We are devising an approach to automatic summarization in the semantic abstraction paradigm, relying on SemRep for semantic interpretation of source text. The transformation stage that condenses these predications is guided by principles articulated in terms of frequency of occurrence as well as lexical semantic phenomena.

We do not produce a textual summary; instead, we present the disorder condensates in graphical format. We first discuss the application of this approach to summarizing single documents (full text research arti-

cles on treatment of disease) and then consider its extension to multidocument input in the form of biomedical scientific abstracts directed at clinical researchers.

The transformation stage takes as input a list of SemRep predications and a seed disorder concept. The output is a conceptual condensate for the input concept. Before transformation begins, predications are subjected to a focused word sense disambiguation filter. Branded drug names such as *Advantage* (Advantage brand of Imidacloprid) and *Direct* (Direct type of resin cement), which are ambiguous with the more common meaning of their names, are resolved to their non-pharmaceutical sense.

3.2 Transformation

In the semantic abstraction paradigm the transformation stage condenses and generalizes, and in our approach these processes are based on four general principles:

a. **Relevance**: Include predications on the topic of the summary

b. **Connectivity**: Also include "useful" additional predications

c. **Novelty**: Do not include predications that the user already knows

d. **Saliency**: Only include the most frequently occurring predications

Although frequency of occurrence (saliency) plays a role in determining predications to be included in the summary, the other three principles depend crucially on lexical semantic information from the UMLS. These four principles guide the phases involved in creating a summary.

Phase 1 (relevance), a condensation process, identifies predications on a given topic (in this study, disorders) and is controlled by a semantic schema (Jacquelinet et al., 2003) for that topic. The schema is represented as a set of predications in which the predicate is drawn from a relation in the UMLS Semantic Network and the arguments are represented as a "domain" covering a class of concepts in the Metathesaurus (Disorders, for example).

{Disorders} ISA {Disorders}
{Etiological process} CAUSES {Disorders}
{Treatment} TREATS {Disorders}
{Body location} LOCATION_OF {Disorders}
{Disorders} OCCURS_IN {Disorders}
{Disorders} CO-OCCURS_WITH {Disorders}

Each domain for the schema is defined in terms of semantic categorization in the Semantic Network. For example {Disorders} is a subset of the semantic group Disorders (McCray et al., 2001) and contains the following semantic types: 'Disease or Syndrome', 'Neoplastic Process', 'Mental or Behavioral Dysfunction', and 'Sign or Symptom'. Although the schema is not complete, it represents a substantial amount of what can be said about disorders. Predications produced by SemRep must conform to this schema in order to be included in the conceptual condensate; such predications are called "core predications."

Phase 2 (connectivity) is a generalization process and identifies predications occurring in neighboring semantic space of the core. This is accomplished by retrieving all the predications that share an argument with one of the core predications. For example, from Naproxen TREATS Osteoarthritis, non-core predications such as Naproxen ISA NSAID are included in the condensate.

Phase 3 (novelty) provides further condensation by eliminating predications that have a generic argument, as determined by hierarchical depth in the Metathesaurus. Arguments occurring less than an empirically determined distance from the root are considered too general to be useful, and predications containing them are eliminated. For example Pharmaceutical Preparations TREATS Migraine is not included in the condensate for migraine because "Pharmaceutical Preparations" was determined to be generic.

Phase 4 (saliency) is the final transformation phase and its operations are adapted from TOPIC's (Hahn and Reimer, 1999) saliency operators. Frequency of occurrence for arguments, predicates, and predications are calculated, and those occurring more frequently than the average are kept in the condensate; others are eliminated.

When these principles are applied to the semantic predications produced by SemRep for a full-text article with 214 sentences (Lisse et al., 2003) concerned with comparing naproxen and rofecoxib for treating osteoarthritis, with respect to effectiveness and gastrointestinal tolerability, the resulting condensate is given in Figure 2. (The abstract for this article was summarized in Figure 1.)

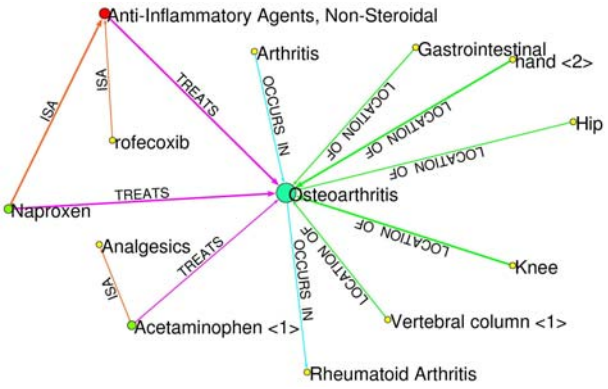

Figure 2. Semantic abstraction summarization of a journal article on osteoarthritis

4 Multidocument Summarization

The MEDLINE database, developed and maintained by the National Library of Medicine, contains more than 12 million citations (dating from the 1960's to the present) drawn from nearly 4,600 journals in the biomedical domain. Access is provided by a statistical information retrieval system. Due to the size of the database, searches often retrieve large numbers of items. For example, the query "diabetes" returns 207,997 citations. Although users can restrict searches by language, date and publication type (as well as specific journals), results can still be large. For example, a query for treatment (only) for diabetes, limited to articles published in 2003 and having an abstract in English finds 3,621 items; limiting this further to articles describing clinical trials still returns 390 citations. We describe the adaptation of our abstraction summarization process to multidocument input for managing the results of searches in MEDLINE.

Extending summarization to multidocument input presents challenges in removing redundancies across documents and at the same time retaining differences that might be important. One issue is devising a framework on which to compute similarities and differences across documents. Radev (2000) defines twenty-four relationships (such as equivalence, subsumption, and contradiction) that might apply at various structural levels across documents. Sub-events (Daniel et al., 2003) and sub-topics (Saggion and Lapalme, 2002) also contribute to the framework used for comparing documents in multidocument summarization.

A particular challenge to multidocument summarization in the extraction paradigm is determining what parts of documents conform to the framework for determining similarities and differences. A recent study (Kan et al., 2001) uses topic composition from text headers, but other studies in the extraction paradigm (Goldstein et al., 1999), extraction coupled with rhetorical structural identification (Teufel and Moens, 2002), and syntactic abstraction paradigms use different methodologies (Barzilay et al., 1999; McKeown et al., 1999).

Our semantic abstraction summarization system naturally extends to multidocument input with no modification from the system designed for single documents. The disorder schema serves as the framework for identifying sub-topics, and predications retrieved across several documents must conform to its structure. Informational equivalence (and redundancy) is computed on this basis. For example, all predications that conform to the schema line {Treatment} TREATS {Disorders} constitute a representation of a subtopic in the disorder domain. Exact matches in this set constitute redundant information, and other types of relationships can be computed on the basis of partial matches. Although we concentrate on similarities across documents, differences could be computed by examining predications that are not shared among citations.

We have begun testing our system applied to the results of MEDLINE searches on disorders, concentrating on the most recent 300 citations retrieved. The results for migraine are represented graphically in Figure 3. Traversing the predicates (arcs) in this condensate provides an informative summary of these citations.

5 Evaluation and Results

Evaluation in automatic summarization, especially for multidocument input, is daunting (Radev et al., 2003). It is usually classified as intrinsic (measures the quality of the summary as related to the source documents) or extrinsic (how the summary affects some other task). Since we do not have a gold standard to compare the final condensates against, we performed a linguistic evaluation on the quality of the condensates generated for four diseases: migraine, angina pectoris, Crohn's disease, and pneumonia. The input for each summary was 300 MEDLINE citations.

Table 1 presents evaluation results. The first author (MF) examined the source sentence that SemRep used to generate each predication and marked the predications as either correct or incorrect. Precision was calculated as the total number of correct predications divided by the total number of predications in the condensate.

We also measured the reduction (compression) for each of the four disorder concepts. In Table 1, "Base" is the number of predications SemRep produced from each set of 300 citations. "Final" is the number of predications left after the final transformation. Therefore, this is a compression ratio on the semantic space of predications, and is different from text compression in the traditional sense.

Concept	Base	Final	C	I	Precision
Migraine	2485	102	72	30	71%
Angina	2989	41	33	8	80%
Crohn's	3077	135	71	64	53%
Pneumonia	2694	28	27	1	96%
Total	11245	306	203	103	66%

Table 1. Results for the four disease concepts
C = Correct, I = Incorrect

In Crohn's disease (with lowest precision) a single SemRep error type in argument identification accounts for 52% of the mistakes. For example in processing the sentence *36 patients with inflammatory bowel disease (11 with ulcerative colitis and 25 with Crohn's disease)*, the parenthesized material caused SemRep to incor-

rectly returned "Inflammatory Bowel Diseases CO-OCCURS_WITH Ulcerative Colitis" and "Ulcerative Colitis predicate CO-OCCURS_WITH Crohn's Disease." Word sense ambiguity also contributed to a large number of errors.

6 Content Characterization

We examined the effect that the transformation stage has on the distribution of predicates and predications during the summarization process. SemRep produced 2,485 predications from 300 citations retrieved for migraine. Of these, 1,638 are distributed over four predicates in the disorder schema (327–TREATS; 148–ISA; 180–LOCATION_OF; 54–CAUSES; 720–OCCURS_IN; and 209–CO-OCCURS_WITH).

After phases 1, 2, and 3 of the transformation process, 311 predications remain (134–TREATS; 41–ISA; 12–LOCATION_OF; 5–CAUSES; 68–OCCURS_IN; and 51–CO-OCCURS_WITH). This reduction is largely due to hierarchical pruning in phase 3.

Phase 4 operations, based on frequency of occurrence pruning (saliency), further condensed the list, and the top three TREATS predication types in the final condensate are (13–Sumatriptan TREATS Migraine; 6–Botulinum Toxins TREATS Migraine; and 6–feverfew extract TREATS Migraine). This list represents the fact that Sumatriptan is a popular treatment for migraine.

Besides frequency, another way of looking at the predications is typicality (Kan et al., 2001), or distribution of predications across citations. Looking at the final condensate for migraine and focusing on TREATS, the most widely distributed predications are "Sumatriptan TREATS Migraine," which occurs in ten citations; "Botulinum Toxins TREATS Migraine" (three citations); and "feverfew extract TREATS Migraine" (two citations).

One can also view the final condensate from the perspective of citations, rather than predications. Of the 300 citations initially parsed, only 63 are represented in the final condensate, one with six predications, one with five predications, three with four predications, and so on. It is tempting to hypothesize that more highly relevant citations will have produced more predications, but this must be formally tested in the context of the user's retrieval objective.

An informal examination of the citations that contributed to the final condensate for migraine revealed differences that we so far do not accommodate. Some of these, such as publication and study type, could be addressed outside of natural language processing with MEDLINE metadata. Others, including medication delivery system and target population of the disorder topic, are amenable to current processing either through extension of the disease schema or enhancements to SemRep.

7 Conclusion and Future Directions

We propose a framework based on semantic abstraction summarization that produces conceptual condensates for disorder topics that are both indicative and informative. The approach uses a biomedical semantic processor as the source interpreter. After semantic interpretation, a series of transformations condense the predications produced, and a final condensate is displayed in graphical form.

In the future, we would like to link the predications in the condensate to the text that produced them. We also plan to evaluate the effectiveness of this approach in retrieving useful articles for clinical researchers. Finally, we would like to investigate additional ways of visualizing the condensates.

Acknowledgements The first author was supported by an appointment to the National Library of Medicine Research Participation Program administered by the Oak Ridge Institute for Science and Education through an inter-agency agreement between the U.S. Department of Energy and the National Library of Medicine.

References

Aronson AR. 2001. Effective mapping of biomedical text to the UMLS Metathesaurus: The MetaMap program. *Proceedings of the AMIA Symp*, pp 17-21.

Batagelj AM. 2003. Pajek - Analysis and Visualization of Large Networks. In M. Jünger and P. Mutzel, editors, *Graph Drawing Software*. Springer Verlag, Berlin, pp 77-103.

Barzilay R, McKeown KR, Elhadad M. 1999. Information fusion in the context of multi-document summarization. *Proceedings of the 37th Annual Meeting of the Association of Computational Linguistics*, pp 550-557.

Burgun A, Bodenreider O. 2001. Comparing terms, concepts, and semantic classes in WordNet and the Unified Medical Language System. *Proceedings of the NAACL Workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations*, pp 77-82.

Clark P, Harrison P, Thompson J. 2003. A knowledge-driven approach to text meaning processing. *Proceedings of the HLT-NAACL Workshop on Text Meaning*, pp 1-6.

Cruse DA. 1986. *Lexical semantics*. Cambridge University Press, Cambridge.

Cutting D, Kupiec J, Pedersen J, Sibun P. 1992. A practical part-of-speech tagger. *Proceedings of the Third Conference on Applied Natural Language Processing*, pp 133-40.

Daniel N, Radev D, Allison T. 2003. Sub-event based multi-document summarization. *Proceedings of HLT-NAACL Workshop on Text Summarization*, pp 9-16.

Fillmore CJ. 1968. The case for case. In E. Bach and RT. Harms, editors, *Universals in Linguistic Theory*. Holt Rinehart and Winston, New York, pp 1-88.

Fellbaum C. 1998. *WordNet*: *An Electronic Lexical Database*. The MIT Press, Cambridge, MA

Goldstein J, Mittal V, Carbonell J, Kantrowitz M. 2000. Multi-document summarization by sentence extraction. *Proceedings of the ANLP/NAACL Workshop on Automatic Summarization*, pp 40-48.

Hahn U, Mani I. 2000. The challenges of automatic summarization. *Computer*, 33(11):29-36.

Hahn U, Reimer U. 1999. Knowledge-based text summarization: salience and generalization operators for knowledge base abstraction. In I. Mani and MT. Maybury, editors, *Advances in Automatic Text Summarization*. MIT Press, Cambridge, pp 215-32.

Harabagiu S, Moldovan D, Pasca M, Mihalcea R, Surdeanu M, Bunescu R; Girju R, Rus V, Morarescu P. 2001. The role of lexico-semantic feedback in open-domain textual question-answering. *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pp 274-81.

Humphreys BL, Lindberg DA, Schoolman HM, Barnett GO. 1998. The Unified Medical Language System: An informatics research collaboration. *J Am Med Inform Assoc*, 5(1):1-11.

Jacquelinet C, Burgun A, Delamarre D, Strang N, Djabbour S, Boutin B, Le Beux P. 2003. Developing the ontological foundations of a terminological system for end-stage diseases, organ failure, dialysis and transplantation. *Int J Med Inf*, 70(2-3):317-28.

Kan M, McKeown KR, Klavans JL. 2001. Domain-specific informative and indicative summarization for information retrieval. *Workshop on Text Summarization (DUC3)*.

Levin B, Rappaport Hovav M. 1996. From lexical semantics to argument realization. *unpublished ms*.

Lisse JR, Perlman M, Johansson G, Shoemaker JR, Schechtman J, Skalky CS, Dixon ME, Polis AB, Mollen AJ, Geba GP. 2003. Gastrointestinal tolerability and effectiveness of rofecoxib versus naproxen in the treatment of osteoarthritis: a randomized, controlled trial. *Ann Intern Med*, 139(7):539-46.

Mani I, Gates B, Bloedorn E. 1999. Improving summaries by revising them. *Proceedings of the 37th Annual Meeting of the Association of Computational Linguistics*, pp 558-65.

McDonald DD. 1992. Robust partial parsing through incremental, multi-algorithm processing. In PS Jacobs, editor, *Text-based Intelligent Systems*. Lawrence Erlbaum Associates, New Jersey, pp 83-99.

McKeown KR, Klavans JL, Hazivassiloglou V, Barzilay R., Eskin E. 1999. Towards multidocument summarization by reformulation: progress and prospects. *Proceedings of the Sixteenth National Conference on Artificial Intelligence*, pp 453-60.

McKeown HR, Chang, SF, Cimino J, Feiner SK, Friedman C, Gravano L, Hatzivassiloglou V, Johnson S, Jordan DA, Klavans JL, Kushniruk A, Pate V, Teufel S. 2001. PERSIVAL, a system for personalized search and summarization over multimedia healthcare information. *JCDL*, pp 331-40.

McCray AT. 1993. Representing biomedical knowledge in the UMLS Semantic Network. High-Performance Medical Libraries: *Advances in Information Management for the Virtual Era*. Meckler Publishing, pp 45-55.

McCray AT, Srinivasan S, Browne AC. 1994. Lexical methods for managing variation in biomedical terminologies. *Proceedings of the Annual Symp Comput Appl Med Care*, pp:235-9.

McCray AT, Burgun A, Bodenreider O. 2001. Aggregating UMLS semantic types for reducing conceptual complexity. *Medinfo*, 10(Pt 1):216-20.

Mihalcea R, Moldovan D. 2000. Semantic indexing using WordNet senses. *Proceedings of the ACL Workshop on IR and NLP*.

Nirenburg S, Raskin V. 1996. Ten choices for lexical semantics. *Memoranda in Computer and Cognitive Science. MCCS-96-304*. New Mexico State University.

Pustejovsky J., Bergler S, Anick P. 1993. Lexical semantic techniques for corpus analysis. *Computational Linguistics*, 19:331-58.

Raphael B. 1968. SIR: Semantic information retrieval. In Minsky, M. (ed.) *Semantic Information Processing*. The MIT Press, Cambridge, pp 33-145.

Radev D. 2000. A Common theory of information fusion from multiple text sources, step one: cross-document structure. *Proceedings of 1st ACL SIGDIAL Workshop on Discourse and Dialogue*.

Radev D, Teufel S, Saggion H, Lam W, Blitzer J, Qi H, Celebi A, Liu D, Drabek E. 2003. Evaluation chal-

lenges in large-scale multi-document summarization: the MEAD project. *Proceedings of ACL*.

Rindflesch TC, Bean CA, Sneiderman CA. 2000. Argument identification for arterial branching predications asserted in cardiac catheterization reports. *Proceedings of the AMIA Symp*, pp 704-8.

Rindflesch TC, Fiszman M. 2003. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. *J Biomed Infor,* 36:462-77.

Saggion H, Lapalme G. Generating indicative-informative summaries with SumUM. 2002. *Computational Linguistics*, 28(4):497-526.

Salton G, Wong A, Yang CS. 1975. A vector space model for automatic indexing. *Communications of the ACM*, (18):613-20.

Srinivasan P, Rindflesch T. 2002. Exploring text mining from MEDLINE. *Proceedings of the AMIA Symp*, pp 722-6.

Sparck Jones K. 1999. Automatic summarizing: factors and directions In I. Mani and MT. Maybury, editors, *Advances in Automatic Text Summarization*. MIT Press, Cambridge, pp 1-13.

Schank RC. 1975. *Conceptual information processing. Amsterdam*. North-Holland Publishing Co, Amsterdam,

Teufel S, Moens M. 2002. Summarizing scientific articles - Experiments with relevance and rhetorical status. *Computational Linguistics*, 28(4):409-445.

Tenny C, Pustejovsk J. 2000. A history of events in linguistic theory. In C. Tenny and J. Pustejovsky, editors, *Events as Grammatical Objects*, CSLI Publications, Stanford, pp 3-37.

Viegas E, Mahesh K, Nirenburg S. 1998. Semantics in action. In P. St. Dizier, editor, *Predicative Forms in Natural Language and in Lexical Knowledge Bases*. Kluwer Academic Publishers, Dordrecht.

Wilks YA. 1976. Parsing English II. In E. Charniak and Y. Wilks, editors, *Computational semantics: An introduction to artificial intelligence and natural language comprehension*. North Holland Publishing Company, Amsterdam, pp 155-84. .

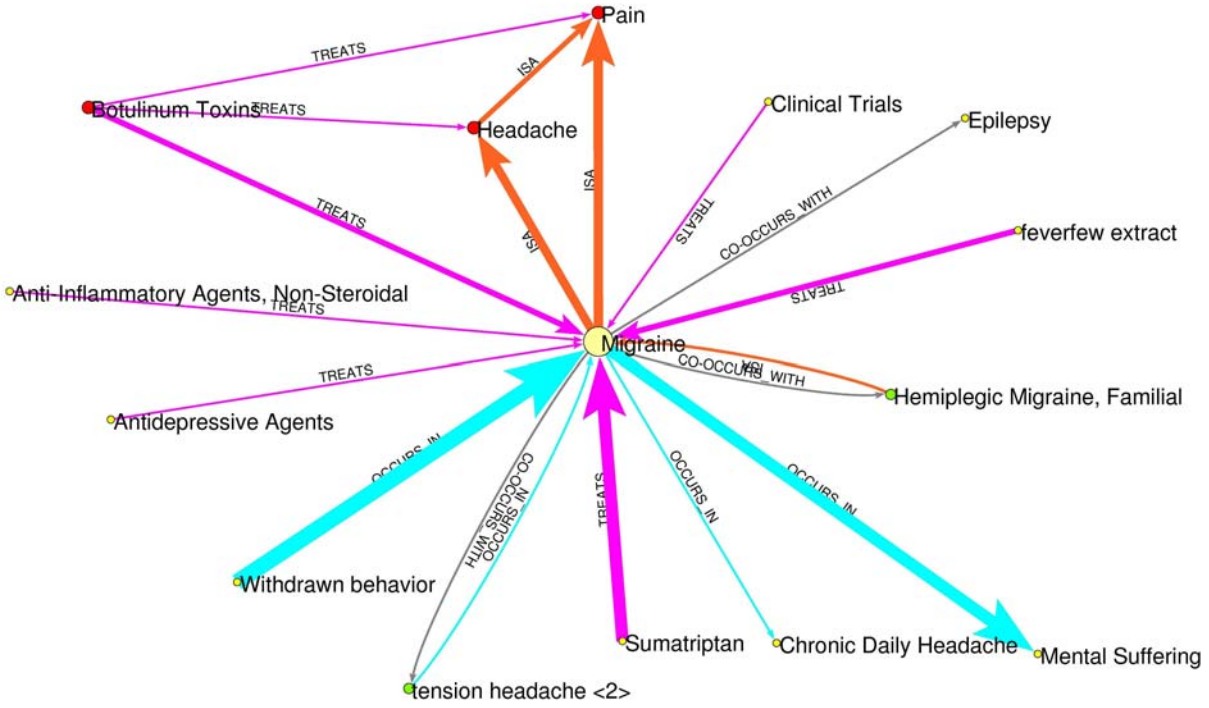

Figure 3. Semantic abstraction summarization on citations retrieved for migraine. Arrow thickness reflects redundant information (i.e. informational equivalence of sentences across multiple documents)