SLIDE 1

The TC project is a collaboration between Susanne Humphrey of CSB and the Lexical Systems Group of CgSB. Humphrey's Journal Descriptor Indexing system is based on words, in particular, characterizing words according to biomedical discipline and high-level category, and therefore it seemed a good fit for developing into a Web-based tool along with the other LexSys tools. And thus the Text Categorization project was born.

The aim of this talk is to give you a basic understanding of the Text Categorization project methodology, mention research on its application to word sense disambiguation, point you to the TC interface so you can try it out yourself, and if time, mention other research.

SLIDE 2

The Text Categorization – or TC – project is concerned with developing tools that categorize text, and also doing research on TC using these tools.

In reality, there are currently two types of categorization in this project, known as:

Journal Descriptor Indexing, or JDI

Semantic Type Indexing, or STI

I'll first be describing JDI and get to STI later.

JDI is concerned with categorizing text according to journal descriptor.

SLIDE 3

What are journal descriptors, or JDs? They are a set of 122 descriptors from the MeSH Vocabulary used for indexing MEDLINE journals per se. JDs are assigned by a human indexer* to the 4100 journals in the training set we use – more about the training set later. The journals and their assigned JDs are part of the List of Serials for Online Users, found in the lsi2007.xml file, which can be ftp'd from the nlmpubs Web site.
*Thanks to Nancy Cox of NLM's Index Section for the intellectual work of maintaining the assignment of JDs for the past several years, and the support of Esther Baldinger of NLM's Serial Records Section.

SLIDE 4

Here we have examples of information from this serials file. The JID, or Journal Unique Identifier; TA,Title Abbreviation; and JDs assigned to three journals – the journal *Transplantation*, with the JD Transplantation; the journal *Pediatric Transplantation*, assigned two JDs Pediatrics and Transplantation, and the *Journal of Pediatric Surgery*, assigned two JDs Pediatrics and Surgery.

SLIDE 5

All 122 JDs are listed, with see and see also references and "includes" notes in the List of Journals Indexed for MEDLINE, also known as the LJI.. LJI contains a Subject Listing where the JDs are headers and the journals are listed under these JD headers.

SLIDE 6

What are the sorts of text that can be JD indexed? To start with, a single word can be indexed, for example the word "transplantation" as shown on this slide. The five top-ranked JDs are shown with their scores, as well as the last-ranked of the 122 JDs. The highest-ranked JD is Transplantation, followed by Hematology, Nephrology, Pulmonary Disease (Specialty), and Gastroenterology. The lowest ranked JD, #122, is Speech-Language Pathology. What this means, in simple terms, is that the word

“transplantation” is found primarily in Transplantation journals – that is journals assigned the JD Transplantation—in our training set (which I’ll get to in a moment), secondarily, the word “transplantation” is found in Hematology journals – that is, journals assigned the JD Hematology, and so forth. The zero score for Speech-Language Pathology means that the word “transplantation” is not found in any of the Speech-Language Pathology journals. The scores are relative to one another; that is, a score has no meaning outside the context of the indexing of this word.

SLIDE 7

What is this training set that I’ve been alluding to? The training set consists of about 3.4 million MEDLINE documents indexed between 1999-2002. JDI requires statistical associations between words in a training set record Title and Abstract and the JDs corresponding to the journal in that training set record. But JDs are not in the MEDLINE record. They are in the NLM serial record from the lsi2007.xml file, I mentioned earlier.

SLIDE 8

As shown here, the JID – Journal Unique Identifier – is in both the training set MEDLINE record, titled “Combined liver and kidney transplantation in children” from the journal *Transplantation*, and also the serial record beneath it for the journal *Transplantation*. This JID serves as the link between the journal cited in a MEDLINE record and the journal in the serial record.

SLIDE 9

In fact, one can think of this link as causing the importation of the JD into the MEDLINE training set record. This slide shows the same training set record titled, “Combined liver and kidney transplantation in children,” with the addition of a JD field containing the value Transplantation. Since the MEDLINE record now has access to the JD of the journal, shown here as imported into the MEDLINE record, we can use co-occurrence

data, specifically the co-occurrence of words in the TI/AB – namely, the words combined, liver, and, kidney, transplantation, children - with the JD Transplantation – in the indexing of text containing these words, as I'll show you in a moment..

SLIDE 10

From now on, I'm going to refer to MEDLINE documents, rather than MEDLINE records, but they are the same thing. So let's go back to our indexing of the word "transplantation", and explain how the score for the top-ranked JD Transplantation is calculated. The score for the JD Transplantation is the number of documents in the training set in which the TI/AB word "transplantation" co-occurs with the JD Transplantation, divided by the number of training set documents in which the word transplantation occurs in the titles/abstracts. The answer must be a number between 0 and 1 – in the case 0.275691.

SLIDE 11

Here we have the Journal Descriptor Indexing of a different word – the word "kidney" – which was also in our training set document - where Nephrology is the highest ranked JD. The Nephrology score 0.140088 is the number of documents in the training set in which the TI/AB word "kidney" co-occurs with the JD Nephrology, divided by the number of training set documents in which the word "kidney" occurs. Each of the approximately 304,000 words in the training set is indexed in this way. This means the system contains all these words with their associated JDs and scores, ready to be used in some way.

SLIDE 12

Now let's consider the indexing of a phrase – specifically the phrase "kidney transplantation". Here are the top-five ranked JDs for this phrase – Transplantation, Nephrology, Hematology, and so forth. A JD score is the average of the JD score for the

word “kidney” and the JD score for the word “transplantation”. Specifically, the score for the top-ranked JD Transplantation, which is 0.178269, is the average of the score for the JD Transplantation, when we indexed the word kidney, and the score for the JD Transplantation, when we indexed the word transplantation. Similarly, the score for the second-ranked JD Nephrology, which is 0.092195, is the average of the score for Nephrology for the word kidney and the score for Nephrology for the word transplantation.

SLIDE 13

And that’s basically how JDI works for JD indexing of a text. The average for a particular JD across the words in the text becomes the score for that JD for the entire text. For example, Nephrology will receive a high score for any text with many “kidney” words in it, such as the phrase shown here - “kidney renal nephron glomerulus”. The particularly strong showing for Nephrology compared to the other JDs is due to the fact that the Nephrology score for each word, when indexed alone, is very high, and therefore the average of these scores must be high as well.

SLIDE 14

It is now possible to perform JD indexing of a document that is outside the training set, such as JDI of the in-process MEDLINE document shown here, based on its title “Kidney transplantation in infants and small children” together with its abstract. The top five JDs are Transplantation, Nephrology, Pediatrics, Hematology, and Urology. The score for each JD is the average of that JD’s scores for words in the title and abstract in this document. Note that for a word to count in a document being indexed, that word must be in our training set.

SLIDE 15

Here is the JD indexing of the same document, but based only on the title, “Kidney transplantation in infants and small children.” Transplantation is the highest ranked JD for both versions. However, in this version, Pediatrics is the second-ranked JD, and Nephrology is the third-ranked JD, where the reverse was the case (Nephrology ranked second, and Pediatrics ranked third) for the title and abstract. Considering the title – “Kidney transplantation in infants and small children” – can anyone guess how the score for Pediatrics was calculated for this title?

SLIDE 16

Answer: The score for Pediatrics is the average of the score for Pediatrics for the words in the title, and most likely Pediatrics has a high score for words like infants and children, and therefore the average for Pediatrics is boosted by these words. The fact that the “native JDs” of the MEDLINE document are Pediatrics and Transplantation – the JDs for the journal *Pediatric Transplantation* – is totally irrelevant. Only words in the title are used for JDI of this title.

SLIDE 17

For example, here is the JDI for a title from *The New England Journal of Medicine*, titled, “Pediatric renal-replacement therapy—coming of age.” returning Nephrology, Pediatrics, and Transplantation as the top three JDs. The native JD for *The New England Journal of Medicine* is Medicine. This example is to emphasize this point – that the native JD of a MEDLINE document being indexed does not at all participate in JD Indexing.

SLIDE 18

Internally, the system has word-JD tables representing the JD indexing of each of the 304,000 words in the training set. The scores for an ordered, such as an alphabetical, list of JDs for a word is also called the word-JD vector for that word. Here is part of the

word-JD vector for the word “kidney” with scores for four of the 122 JDs – Nephrology, Psychiatry, Psychopharmacology, and Transplantation – in alphabetic order. Note the scores for Nephrology and Transplantation are relatively high, compared to the scores for Psychiatry and Psychopharmacology.

SLIDE 19

Here is part of the word-JD vector for the word “renal” showing scores for the same JDs. Again, the scores for Nephrology and Transplantation are relatively high, compared to those for Psychiatry and Psychopharmacology.

SLIDE 20

Now we show the word-JD vector for the word “schizophrenia”. Unlike kidney and renal, the scores for Psychiatry and Psychopharmacology are relatively high, compared to those for Nephrology and Transplantation. The zero score for Nephrology is because the word schizophrenia does not appear in any Nephrology journal in the training set..

SLIDE 21

The importance of this, is that there are standard measures comparing JD vectors to one another resulting in similarity scores between 0 and 1. The similarity of the JD vector for the word kidney compared to itself is 1.0. The similarity of the JD vector for the word kidney and the JD vector for the word renal is 0.96. But the similarity of the JD vector for the word kidney and schizophrenia is 0.03. The measure we use in our project is the vector cosine coefficient from the well-known textbook by Salton and McGill.

SLIDE 22

The next three slides show the vector cosine coefficient formula, first for calculating the similarity between the JD vectors of any two words, WORD-i and WORD-j

SLIDE 23

the similarity between the JD vector of any word and the JD vector of any document, WORD-i and DOC-j.

SLIDE 24

and finally the similarity between the JD vectors of any two documents, DOC-i and DOC-j.

SLIDE 25

There are two avenues of research involving JD vector similarity which we are contemplating – one involving pairs of words in the training set, and another that compares word-JD vectors to document-JD vectors.

An example of research involving similarity between word-JD vectors is the automatic creation of stopword lists; our current stopword list was developed empirically. Here, we compare the JD vector for the quintessential stopword THE (where the 122 JD scores range from 0.012152 down to 0.000048) to all the other words in the training set. In theory, a word with a JD vector similar to THE – with all low, gradually decreasing scores) would likely be a good stopword as well.

Another avenue of research involves comparing JD vectors of different words to the same MEDLINE document. In theory, the more similar a word JD vector is, to a MEDLINE document JD vector, the more descriptive that word is of the MEDLINE document, and by contrast, a JD vector for a word that is very dissimilar to a MEDLINE document JD vector, would not be a good descriptor for the document. Thus, an indexing term assigned to a document – whether as a recommendation from an automated indexing system such as MTI or humanly-assigned – might be detected as an outlier because of the

great dissimilarity of the term's JD vector to the JD vector of the document being indexed.

SLIDE 26

Here we show a result of similarity of the word-JD vector for the word THE to its most similar words. Similarity of THE to itself is, of course 1.0. To AND is 0.9998. To FOR is 0.9977. To WITH is 0.9970. The most dissimilar word in the training set is COMLEX (which is an acronym for Comprehensive Osteopathic Medicine Licensing Examination, and is exclusively associated with the JD Osteopathy).

SLIDE 27

The impetus for the outlier detector was the coming to our attention that MTI was recommending the indexing term Stupor resulting from the word "unresponsive" in a MEDLINE document even when the document was referring to unresponsive cells. This recommendation would be considered a blooper for indexing such a document. As shown in this slide, the similarity between the JD vector for the term "Stupor" and for the title/abstract of MEDLINE document titled, Human intestinal epithelial cells are broadly unresponsive to Toll-like receptor2-dependent bacterial ligands: implications for host-microbial interactions in the gut, is only about .2. However, the similarity between the other indexing terms is much higher, for example, 0.9 for the recommendation Toll-Like Receptor 2. So, Stupor stands out as an inappropriate indexing term compared to the others, based on vector similarity. We are investigating if this phenomenon can be used for detecting other such blooper recommendations.

Bloopers can also occur in human indexing. For example, I recently came across a MEDLINE document (PMID 12965020) indexed metaphorically (and therefore incorrectly) under the MH Deception (a social behavior term), on account of the notion of cheating, as in the title "Competitive fates of bacterial social parasites: persistence and self-induced extinction of Myxococcus xanthus cheaters." A successful blooper detector

would compute a low similarity of 0.14 between the JD vector for the term Deception and the JD vector for the TI/AB of this document, compared to a high similarity of 0.82 between the JD vector for the term “Myxococcus xanthus” and the document.

JDI of the MH Deception:
1|0.229599|Psychology
2|0.057222|Psychiatry
3|0.054264|Behavior
4|0.046512|Jurisprudence
5|0.038760|Ethics

JDI of the word Deception
1|0.220074|Psychology
2|0.088802|Psychiatry
3|0.063158|Psychophysiology
4|0.052632|Behavior
5|0.042105|Nursing

JDI for PMID 12965020
TI - Competitive fates of bacterial social parasites: persistence and self-induced extinction of Myxococcus Xanthus cheaters.
1|0.075398|Microbiology
2|0.066032|Bacteriology
3|0.033764|Science
4|0.029634|Molecular Biology
5|0.029304|Biology

In fact, the similarity between the JD vector for the MH Deception and the title of this MEDLINE document is quite low – 0.12. Similarity between the JD vector for the word “deception” and the title of this MEDLINE document is quite low – 0.14. By contrast, similarity between JD vector for MH Myxococcus xanthus under which this document is

indexed and title is 0.82, and similarity between JD vector for phrase “Myxococcus Xanthus” and title is 0.84.

SLIDE 28

Now let’s talk briefly about Semantic Type Indexing. Semantic types are the set of 135 Semantic Types in the Semantic Network in NLM’s UMLS (Unified Medical Language System). Concepts in the UMLS Metathesaurus are assigned one or more STs which semantically characterize those concepts. For example, the concept “aspirin” is assigned the STs Pharmacologic Substance (phsu) and Organic Chemical (orch)

SLIDE 29

Just as the system contains word-JD tables representing JD indexing for each training set word, the system also contains word-ST tables representing the semantic type indexing of each training set word. There isn’t time to explain in detail how these tables are built, but in general, it is based on JDI and computing JD vector similarity. Thus, a text can be indexed according to ST, just as it can be indexed according to JD.

SLIDE 30

Research has been published on ST indexing as a tool for disambiguating text. Disambiguation is a major challenge in natural language processing, such as that performed by MetaMap, on which the automated Medical Text Indexer is based.
STI was used for disambiguating 45 ambiguous strings from NLM’s WSD collection, which had been disambiguated by humans as the gold standard. The number of instances for each ambiguity ranged from 3 to 67, with an average of 54. Instances for which “None of the Above” was the gold standard were ignored, since neither STI nor the baseline method to which it was compared was designed to return this answer. The study was published in January 1, 2006, issue of JASIST (Humphrey SM, Rogers WJ, Kilicoglu H, Demner-Fushman D, and Rindflesch TC. Word sense disambiguation by

selecting the best semantic type based on Journal Descriptor Indexing: preliminary experiment. *Journal of the American Society for Information Science and Technology*. 2006 Jan 1;57(1):96-113. Erratum in: *J Am Soc Inf Sci Technol* 2006 Mar;57(5):726.)

SLIDE 31

For example, the ambiguity "transport" has two meanings: "Biological Transport" assigned the ST Cell Function (celf) and "patient transport" assigned the ST Health Care Activity (hlca). The STI methodology can analyze text, such as a MEDLINE document, containing an ambiguous string and determine which of the STs assigned to that string by UMLS receives a higher score for that text, which then returns the associated meaning, presumed to apply to the ambiguity itself.

SLIDE 32

For example, the input corresponding to title and abstract of PMID 9674486 contains the ambiguity "transporting" (a variant of transport) in the last sentence, "This practice averts the potential complications associated with transporting critically ill patients." When a system like MetaMap encounters such an ambiguity, it needs to know the correct meaning. We as humans can easily disambiguate the word "transporting", choosing the correct ST of hlca (for Health Care Activity) over the ST celf (for Cell Function). Automatic STI also successfully performed this disambiguation, according to the higher score for the ST hlca for this document, compared to celf.

SLIDE 33

One of the issues is the context of the ambiguity, may be the one sentence with the ambiguity, all sentences with the ambiguity, or the entire MEDLINE document.
In this study, STI achieved an overall average precision of 0.7873 compared to 0.2492 for a baseline method known as MeSH Frequency. The baseline method involves automatically matching each candidate concept for an ambiguity to a MeSH synonym if

there is one. The concept matching the MeSH synonym with the highest frequency count in MEDLINE is returned as the answer. If some concept has no MeSH synonym, then it has no chance of being the answer. For example, if there are two candidates for an ambiguity, and the correct one has no corresponding MeSH synonym, then the other concept wins for all instances of the ambiguity in the collection, even if the first candidate is the correct answer for most or even all the instances.
STI continues to be investigated for WSD in NLP applications related to the Indexing Initiative and Semantic Knowledge Representation.

SLIDE 34

Most of the JDI and STI in this talk can be done by using the TC Web Tools at the TC Web site. TC tools and applications are freely distributed with open source code, 100% in JAVA, running on different platforms. It is one complete package with documentation and support, and provides Java APIs and command line tools. We are in our first release of TC 2007.
You can click on Documentation at the TC Web site for links to our publications, including the WSD paper. In coming months, we will be adding to the functionality of TC Web tools as well as incorporate the ability to create new training sets.
The JAVA system was developed by Chris Lu in the Lexical Systems Group and authorized by Allen Browne, both of the Lexical Systems Group; Willie Rogers, working under the Indexing Initiative in CgSB, is a collaborator on the TC project.

SLIDE 35

JDI has been used for several years by SemRep as a pre-processing step to increase accuracy by identifying MEDLINE documents in the molecular genetics domain before NLP begins.

We also have some ideas on research involving JDI. Here are some of them listed:

Evaluating JDI. One approach would be to take a random sample of recent MEDLINE documents, JDI them, and use as a criterion of success whether the native JD of the document (which doesn't participate in JDI) was ranked highly in the JDI result.

Creating specialty subsets of general medical journals, such as The New England Journal of Medicine or JAMA, or the journal Science. Or partitioning any large, varied collection into specialties for users who would like to be alerted to relevant material in their specialty or some intersection of specialties.

JDI is word-based. Some have suggested that it be phrase-based, or that we consider variants of a word to be a single word.

One could possibly expand JDI beyond biomedicine by using LC call numbers as JDs, and developing a training set from collections representing all subjects.

SLIDE 36

Just to illustrate creating specialty subsets. There's a real-world example on the Web site of the American Academy of Pediatrics. Editors have been categorizing published studies in the journal *Pediatrics*, since January 1997, according to subspecialties similar to JDs. For example, a pediatric oncologist can select Tumors as a link to full-text articles in this journal on the subject of childhood cancer, beginning with the most recent issue of the journal and going back in time.

SLIDE 37

Another real-world example is the Web site of the journal Science, published by the American Association for the Advancement of Science. Since 1996 editors have been categorizing published studies in the journal Science, according to Science Subject Collections. For example, a meteorologist can select Atmospheric Science as a link to

articles in this journal on this subject, ordered by most recently published. There is also a search box for entering keywords in a selected Collection..

SLIDE 38

I don't have time to go into this now, but there are also MH-JD vectors from the training set – that is, the JD vectors of indexing terms in the training set. In theory, one can use the vector similarity computation to perform automated MeSH indexing of an unindexed document. This would involve comparing each of the 20,000-plus MH-JD vectors from the training set against the TI/AB-JD vector of the MEDLINE document being indexed, and then ranking the MHs in terms of similarity of their JD vectors to this TI/AB-JD vector.

In general, whenever you have an X-JD vector and Y-JD vectors, you can create an X-Y vector, based on comparing the similarity of each Y-JD vector to the X-JD vector. In this case, one has a word-JD vector for a word in a document and MH-JD vectors, and can create a word-MH vector. Then, the score of an MH for the document would be the average of that MHs score for the words in the document.