Automatic Summary Creation by Applying Natural Language Processing on Unstructured Medical Records

Daniela Giordano, Isaak Kavasidis(✉), and Concetto Spampinato

Department of Electrical, Electronics and Informatics Engineering,
University of Catania, Viale Andrea Doria, 6, 95125 Catania, Italy
{dgiordan,ikavasidis,cspampin}@dieei.unict.it

Abstract. In this paper we present a system for automatic generation of summaries of patients' unstructured medical reports. The system employs Natural Language Processing techniques in order to determine the most interesting points and uses the MetaMap module for recognizing the medical concepts in a medical report. Afterwards the sentences that do not contain interesting concepts are removed and a summary is generated which contains URL links to the Linked Life Data pages of the identified medical concepts, enabling both medical doctors and patients to further explore what is reported in. Such integration also allows the tool to interface with other semantic web-based applications. The performance of the tool were also evaluated, achieving remarkable results in sentence identification, polarity detection and concept recognition. Moreover, the accuracy of the generated summaries was evaluated by five medical doctors, proving that the summaries keep the same relevant information as the medical reports, despite being much more concise.

1 Introduction

Every day a large amount of medical reports, in the form of free text (i.e. not structured according to a logical scheme) is generated. Not possessing any structural information hampers the ability of automatic document digitization and analysis and subsequently all the applications that could be built upon these. The information included in the text can be deductible only through reading. The adoption of free text documents is done mainly due to the doctors' lack of time, who have to write reports quickly, or due to hospitals' internal procedures or traditions. Moreover, the readability of these documents could become a problem as it may not be easy for the reader to pinpoint the most important parts.

The medical domain suffers particularly by an overload of information and rapid access to key information is of crucial importance to health professionals for decision making. For instance, a concise and synthetic representation of medical reports (i.e. a summary), could serve to create a precise list of what was performed by the health organization and derive an automatic method for

© Springer International Publishing Switzerland 2015
G. Azzopardi and N. Petkov (Eds.): CAIP 2015, Part II, LNCS 9257, pp. 383–393, 2015.
DOI: 10.1007/978-3-319-23117-4_33

calculating hospitalization costs. Given the plethora in number and diversity of sources of medical documents, the purpose of summarization is to make users able to assimilate and easily determine the contents of a document, and then quickly determine the key points of it. In particular, as reported in [1]: *"A summary can be loosely defined as a text that is produced from one or more texts, that conveys important information in the original text(s), and that is no longer than half of the original text(s) and usually significantly less than that"*, but also denotes its most important challenge: *"Identifying the information segments at the expense of the rest is the main challenge in summarization"*. Generating summaries, however, is not trivial as it implies a deep understanding of the underlying semantics. This is even more challenging in the medical domain since medical reports include a highly specialized vocabulary, words in upper and lowercase letters and numbers that require ad-hoc tokenization. These problems urged the development of domain-specific resources such as PubMed/MEDLINE and PubMedCentral[1], ontologies and other semantic lexical resources, such as Gene Ontology[2] and Unified Medical Language System *(UMLS)*[3], and annotated databases, such as Entrez Gene[4] which are used heavily by a variety of text mining applications.

The objectives of the work presented herein is 1) to create automatically a summary that conveys the key points of medical reports and 2) to provide a tool for annotating the medical concepts found in the text with Linked Life Data (*LLD*)[5], so that the doctors or the patients can explore further what is being reported and also enable interoperability with other semantic web-enabled applications.

The remainder of the paper is as follows: the next section briefly presents related works, while Section 3 describes the method in detail and in Section 4 a performance evaluation of the system is carried out. Finally, in the last section conclusions are drawn and future works are given.

2 Related Work

Text summarization of medical documents was brought to the attention of the scientific community due to the tremendous growth of information that are available to physicians and researchers: the growing number of published journals, conference proceedings, medical sites and portals on the World Wide Web, electronic medical records, etc.

In particular, in the clinical context, there has been an increase of interest in the use of Electronic Medical Records *(EMR)* systems which may contain large amounts of text data, to improve the quality of healthcare [14]. To make full use of the information contained in the EMR and to support clinical decision,

[1] http://www.ncbi.nlm.nih.gov/pubmed
[2] http://www.geneontology.org/
[3] http://www.nlm.nih.gov/research/umls/
[4] http://www.ncbi.nlm.nih.gov/gene/
[5] http://linkedlifedata.com/

text mining techniques based on Natural Language Processing *(NLP)* have been especially proposed for information retrieval purposes or for extracting clinical summaries.

In [16], an information extraction system that extracts three types of information (numeric values, medical terms and categories) from semi-structured patient records, is presented. An extension to this system is presented in [15]: The MEDical Information Extraction (MedIE) system extracts a variety of information from free-text clinical records of patients with breast related diseases. MedIE uses GATE [5], WordNet [11] and UMLS, and employs a graph-based approach for numeric attribute extraction capable of performing the majority of information extraction tasks achieving remarkable results. In [10], the Keyphrase Identification Program *(KIP)* is proposed, for identifying medical concepts in medical documents. KIP combines two functions: noun phrase extraction and keyphrase identification. It automatically extracts phrases containing nouns using a part-of-speech tagger achieving fair results (0.26 in precision and 0.60 in recall, best case scenario). KIP ranks all the noun phrases in terms of their relevance to the main subject of the document, and selects only the most relevant ones by creating a glossary database from the Medical Subject Headings (MeSH) site. In [12] is presented a pipeline-based system for automated annotation of surgical pathology reports with UMLS terms built on GATE. The system implements a simple method for detecting and annotating UMLS concepts as well as annotating negations based on the NegEx algorithm [4], achieving very good results in terms of precision (0.84) and recall (0.80). In [13] another example of application that mines textual information by employing NLP methods is presented, but this time such information is being integrated with other types of biological data found on-line.

While all of these tools offer great insight on how concept identification and annotation can be done they do not offer any functionalities for single-document text summarization. Such feature can be found in more complex works, as in [9,2] where summarization of single documents is done by applying robust NLP techniques combined with conceptual mapping based on ad-hoc ontologies or lexicons. The main problem with these approaches is that the accuracy of the concept extraction, and subsequently the accuracy of the summarization, depends on the underlying lexicon, and in this particular case, the ontology. Not using well established ontologies carries the drawback of limiting the available identifiable concepts and also, their interoperability with other semantic web-based complementary systems. In [8], *UMLS* is used for concept mapping but the system does not deal with negative expressions leading to misinterpretations in the final summary.

In the next section the description of a system aiming at creating summaries out of medical records written in free text form by implementing a GATE pipeline, and also for assigning UMLS codes to the medical entities found inside them, is proposed.

3 Method

In order to produce a reliable summary, the corpus of medical documents must undergo through several processing steps. In this section, the tools used during this process are introduced and described. The basis of the developed system is GATE, which is the most used tool for implementing NLP-based applications. GATE uses regular expressions to configure all of its components (Tokenization, Sentence Splitter, POS tagging, Named Entity Recognition (NER) etc...).

The general architecture of the proposed system is shown in Fig. 1.

3.1 Text Processing and Annotation

ANNIE [6] is the information extraction component of the GATE platform and it substantially encapsulates the main NLP functions. In our case, an ANNIE pipeline was defined that employs the following components:

- **English Tokenizer:** The text in the corpus is divided into very simple tokens such as numbers, punctuation symbols or simple words. The main objective of this module is to maximize the efficiency and flexibility of the whole process by reducing the complexity introduced by the grammar rules.
- **Gazetteer:** Its role is to identify the names of entities based on lists, fed into the system in the form of plain text files. Each list is a collection of names, such as names of cities, organizations, days of the week, etc...
- **Sentence Splitter:** As its name suggests, it splits the text in simple sentences by using a list of abbreviations to distinguish sentence markers.
- **Part-of-speech Tagger:** Marks a word as corresponding to a particular part of speech based on both its definition and context. This is useful for the identification of words as nouns, verbs, adjectives, adverbs, etc. The results of this plug-in are the tokens used for the implementation of regular expressions.
- **Named Entity Transducer:** ANNIE's semantic tagger contains rules that work on the annotations of the previous phases to produce new annotations. It is used to create annotations regarding the terms related on negations, sections and phrases.
- **MetaMap Annotator:** This module serves the role of identifying medical terms found in text and map them to UMLS concepts by using NLP methods combined with computational linguistics [3].
- **Words Correction:** Given that the vast majority of the medical reports that we are dealing with were produced in a completely manual manner, misspellings do occur, making the medical term identification process less accurate. For this reason, each unannotated term (i.e. a word that does not exist) in the text is used as a query term against a dataset containing medical terms and the term with the smallest Levenshtein distance is retrieved. The result is used in place of the misspelled word in the original document.
- **Negated Expressions:** In order to achieve a correct interpretation of the text found in medical documents, it is very important be able to identify

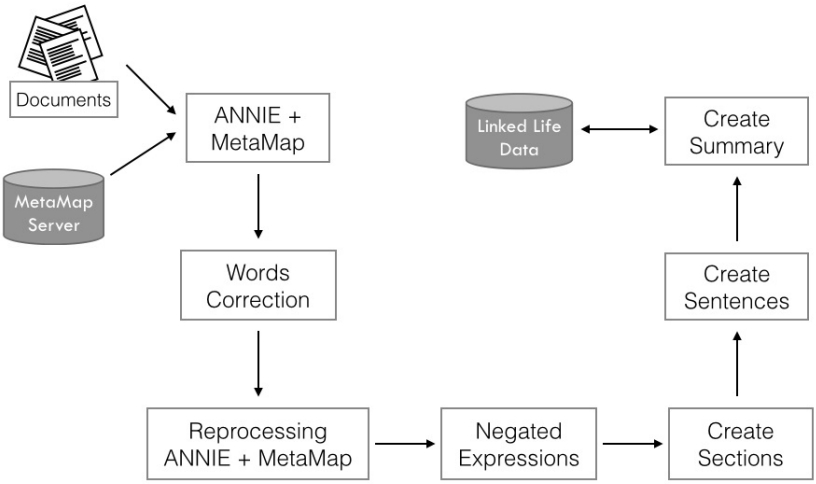

Fig. 1. General architecture of the proposed system.

negated expressions, which indicate the absence of a particular symptom or condition. MetaMap helps to identify negated concepts by providing a pair of features, namely "NegExType" and "NegExTrigger"; the former one identifies the negation, while the latter one specifies the term that expresses it. In this phase there are two problems that must be dealt with: a) the negated medical concept must be correlated to the term that triggers the negation effect and b) there are words that imply negation but MetaMap cannot identify them as such (e.g. the word inexistence). To overcome these problems, the Gazetteer is used again, by creating a new class of annotations relating exclusively to terms of negation.

– **Section Parsing:** For this phase, the Gazetteer plug-in is used by defining tags that could be possibly represent section labels. For our experiments the following tags were defined: admitting diagnosis, discharge diagnosis, symptoms, past medical history, family history, social history, hospital course, medications, diagnostic studies, discharge instructions.

3.2 Summary Generation

Not all of the annotations generated by the MetaMap Annotator are needed in the final summary. Each MetaMap annotation contains also the semantic type of the corresponding term (e.g. "Body Part" for the word "leg", "Manufactured Object" for the word "scalpel" etc...). Inevitably, terms belonging to certain semantic types are excluded from the summary because their importance might be negligible.

An issue that needs to be dealt with during summary generation is that many annotated phrases should be merged to one sentence. For example, the sentence "x-rays including left foot, right knee, left shoulder and cervical spine" would normally be divided in the tokens "x-rays", "left foot", "right knee", "cervical spine" and "left shoulder" even though all of them belong to the same sentence.

Regular expressions were employed to face this problem. In our case, the following regular expression was used:

$$(PRE)?(NEG)?((METAMAP)(NEG)?)+(POSTCONCEPT)?(POST)?,$$

where $METAMAP$ denotes the main medical concept identified by MetaMap (e.g. "amoxicillin", PRE denotes attributes that can precede the main concept (e.g. "significant", "treated with", "diagnosis of", "presence of" etc...), $POSTCONCEPT$ indicates a word directly correlated to the main concept (e.g. "1 g" for expressing dosage etc...) and $POST$ denotes eventual tokens that may represent a continuation of the sentence (e.g. commas, conjunctions etc...). Finally, the NEG term indicates whether a token expresses negativity or not.

The "+" and "?" operators describe the cardinality of each term with the "+" operator meaning "at least one or more" and the "?" operator meaning "zero or more".

For each identified section, the annotations relative to affirmative and negative expressions are created and for each sentence, the annotations produced by MetaMap are used. The same annotations are also used as query terms on the LLD site and the URLs pointing to the corresponding medical concepts are embedded to the final summary and exported in an HTML file.

An example of how the system works is shown below. Given the following discharge summary (the underlined words represent typographical errors):

ADMITTING DIAGNOSES: Intrauterine pregnancy at 36 weeks. Twin gestation. Breech presentation of twin A.
DISCHARGE DIAGNOSES: Intrauterine prengancy at 36 weeks. Twin gestation. Breech presentation of twin A. Status post primary low transverse cesarean section for malpresentation of twins.
CHIEF COMPLAINT: At the time of admission, contractions.
HISTORY: The patient is a 32-year-old pregnant at 36 weeks with known twins with contractions and good fetal movement, no bleeding, no loss of fluids.
OB HISTORY: Present pregnancy with previous receipt of a steroid window.
GYN HISTORY: Significant for chamydia, which was treated.
MEDICATIONS: Prenatal vitamins.
SOCIAL HISTORY: No drinking, smoking or drug use. No domestic violence. The father of the baby is currently involved, and the patient is living with a friend.
PHYSICAL EXAMINATION: Temperature is 36.2, pulse 88, respirations 18 and blood pressure 121/58. HEART: Regular rate and rhythm. LUNGS: Clear. ABDOMEN: Soft and gravid.
HOSPITAL COURSE: Postoperatively, the patient did well. She was eating, ambulating and voiding, passing gas by postoperative day 2, and on postoperative day 3, she continued to do well. She had been seen by Social Work and

options made aware to the patient. She was ready for discharge. She remained afebrile throughout her hospital course.
DISCHARGE INSTRUCTIONS: She will be discharged to home to follow up in two weeks for a wound check.
MEDICATIONS AT THE TIME OF DISCHARGE: Percocet, Motrin and Colace.

The result is a more compact form of the input document, with both the wrong words corrected and also contains the Linked Life Data links identified by MetaMap:

ADMITTING DIAGNOSIS: Intrauterine pregnancy. Breech presentation of twin.
SYMPTOMS: contractions.
DISCHARGE DIAGNOSIS: Intrauterine pregnancy. Breech presentation of twin. Malpresentation of twins.
DIAGNOSTIC STUDIES: Temperature 36.2, pulse 88, respirations 18 and blood pressure 121/58. HEART. LUNGS. ABDOMEN.VAGINAL
PAST MEDICAL HISTORY : Significant for chlamydia. known twins with contractions and good fetal movement ,. pregnancy. Receipt of a steroid window.
PAST MEDICAL HISTORY NEGATIVE: no bleeding, no loss of fluids.
SOCIAL HISTORY NEGATIVE : No drinking, smoking or drug use. No domestic violence.
MEDICATIONS : Prenatal vitamins. Percocet, Motrin and Colace.

By clicking on the underlined terms, the system redirects the reader to its *LLD* page (Fig. 2).

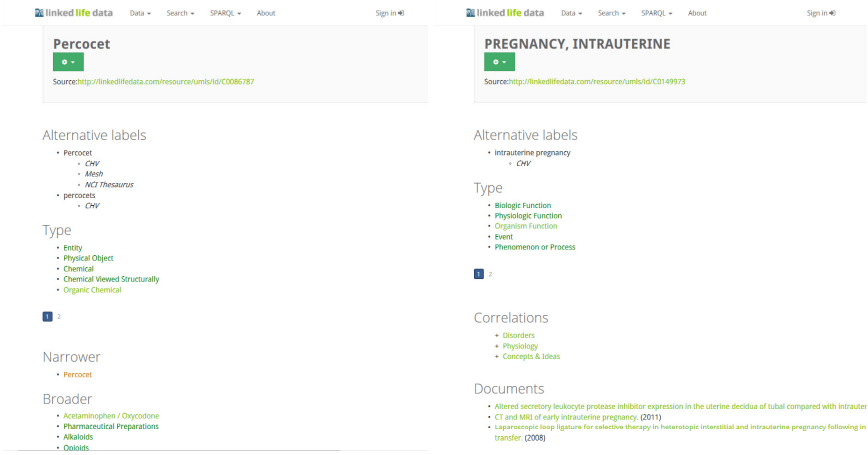

Fig. 2. Image showing the *LLD* pages of the terms *Percocet* (left) and *Intrauterine pregnancy* (right)

4 Performance Evaluation

As stated in [1], evaluating the performance of a summarization system is not a trivial task. To be more precise, while the quantitative evaluation can be based on clear and objective metrics, the qualitative one is not that straightforward because summarization efficiency is most often expressed as a subjective opinion of the individual rater (i.e. Inter-rater reliability). Nevertheless, because of the two-fold nature of these kind of systems, their performance evaluation should cover both these aspects. So, in order to assess exhaustively the performance of the proposed system we tested it under three different perspectives and compared the results to a hand-crafted ground-truth (described in Subsection 4.1). For all the evaluations we employed Precision-Recall and F_1 measure values defined as follows:

$$Precision = \frac{TP}{TP + FP},$$

$$Recall = \frac{TP}{TP + FN}$$

and

$$F_1 = \frac{Precision \times Recall}{Precision + Recall}$$

The FP, TP and FN values are defined separately for each of the aspects tested. The obtained results were compared against a manually created dataset by five medical doctors that contained both positive and negative sentences. The dataset was comprised by 125 medical reports containing 3611 annotated sentences (2824 positive and 787 negative) and 15641 annotated medical concepts.

- **Medical concept recognition**: The first aspect of the system that was tested was its ability to identify correctly the medical concepts found inside the medical reports.
  - A True Positive *(TP)* results when an identified medical concept is the same with the manual annotation.
  - A False Negative *(FN)* results when a medical concept was not identified correctly or was not identified at all.
  - A False Positive *(FP)* results when a medical concept was assigned a different label or when a non medical term was identified as such.

Table 1. Performance of the system in recognizing correctly the medical concepts.

N	TP	FP	FN	P	R	F_1
15641	12499	2419	3142	0.84	0.8	0.82

– **Sentence identification and polarity detection**: The second aspect of the system that was tested was its ability to extract correctly the single sentences in the medical report and also to assign correctly the negation attribute to the medical concepts detected by the previous test, using regular expressions.
  - A True Positive *(TP)* results when an identified sentence is found also in the ground truth and was assigned the correct polarity.
  - A False Negative *(FN)* results when a sentence found in the ground truth was not identified as such or when an annotated sentence was divided erroneously between two other sentences or when the negation property was not assigned to a negative sentence .
  - A False Positive *(FP)* when a sentence is erroneously identified as such, but instead, in the ground truth, its terms do not belong in the same one or when the negation property was assigned to a positive sentence.

Table 2. Performance of the system on sentence detection and polarity detection.

N	TP	FP	FN	P	R	F_1
3611	2808	531	803	0.84	0.78	0.81

– **Summary relevance**: Additionally, the quality of the produced summary was evaluated. To achieve this, the same five medical doctors were presented with both the original reports and the final results and then asked to assess qualitatively the relevance of the summaries (i.e. express their personal opinions on what medical concepts should be included in the final summary versus what should be excluded). After that, the following parameters were defined:
  - A True Positive *(TP)*: A concept that the medical doctors felt that should be included in the final summary and it was.
  - A False Negative *(FN)*: A concept that the medical doctors felt that should be included in the final summary but it was not.
  - A False Positive *(FP)*: A concept that the medical doctors felt that should not be included in the final summary but it was.

Table 3. Performance of the system on summary accuracy. The final result was calculated based on the sum of the votes of the medical doctors.

N	TP	FP	FN	P	R	F_1
15641	11499	3514	4142	0.77	0.74	0.75

5 Discussion

Sentence identification and polarity detection performance was very good. Indeed, an $F_1 score$ value of 0.81 means that the algorithms employed to do this task performed very well. More detailed inspection of the failing sentences were due to misplaced punctuation marks and missing negative keywords from the employed dictionary that could provoke ambiguity problems if they were ultimately included (e.g. the word *"will"* in the sentence *"...will develop cancer..."* does not imply that the patient has cancer). The results in medical concept recognition are almost equal as high. An $F_1 score$ value of 0.82 means that the MetaMap module is very accurate in identifying the medical concepts found in the reports. Especially important are the results in the summary accuracy test where the subjective opinion of the intended end users of the system (the medical doctors) determine its utility. An $F_1 score$ value of 0.75 implies that the generated summaries are valid and also demonstrates that the proposed system can be a robust solution for other applications that will make use of its functionalities, such as [7].

In this paper a system that automatically generates summaries taking as input the corpus of unstructured medical reports, was presented. Such summaries, are also annotated with links which the reader can follow in order to get a short description of the corresponding medical concepts. The same system could be configured to use the International Classification of Diseases (ICD) dictionary, instead of or in addition to UMLS, to assign codes to diseases making the system more compatible with existing systems.

References

1. Afantenos, S., Karkaletsis, V., Stamatopoulos, P.: Summarization from medical documents: a survey. Artificial Intelligence in Medicine **33**(2), 157–177 (2005)
2. Aramaki, E., Miura, Y., Tonoike, M., Ohkuma, T., Mashuichi, H., Ohe, K.: Text2table: medical text summarization system based on named entity recognition and modality identification. In: Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing, pp. 185–192. Association for Computational Linguistics (2009)
3. Aronson, A.R.: Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In: Proceedings of the AMIA Symposium, p. 17. American Medical Informatics Association (2001)
4. Chapman, W.W., Bridewell, W., Hanbury, P., Cooper, G.F., Buchanan, B.G.: A simple algorithm for identifying negated findings and diseases in discharge summaries. Journal of biomedical informatics **34**(5), 301–310 (2001)
5. Cunningham, H.: Gate, a general architecture for text engineering. Computers and the Humanities **36**(2), 223–254 (2002)
6. Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V.: GATE: a framework and graphical development environment for robust NLP tools and applications. In: Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL 2002) (2002)

7. Giordano, D., Kavasidis, I., Spampinato, C., Bella, R., Pennisi, G., Pennisi, M.: An integrated computer-controlled system for assisting researchers in cortical excitability studies by using transcranial magnetic stimulation. Computer methods and programs in biomedicine **107**(1), 4–15 (2012)
8. Johnson, D.B., Zou, Q., Dionisio, J.D., Liu, V.Z., Chu, W.W.: Modeling medical content for automated summarization. Annals of the New York Academy of Sciences **980**(1), 247–258 (2002)
9. Lenci, A., Bartolini, R., Calzolari, N., Agua, A., Busemann, S., Cartier, E., Chevreau, K., Coch, J.: Multilingual summarization by integrating linguistic resources in the mlis-musi project. LREC **2**, 1464–1471 (2002)
10. Li, Q., Wu, Y.F.B.: Identifying important concepts from medical documents. Journal of biomedical informatics **39**(6), 668–679 (2006)
11. Miller, G.A.: Wordnet: a lexical database for english. Communications of the ACM **38**(11), 39–41 (1995)
12. Mitchell, K.J., Becich, M.J., Berman, J.J., Chapman, W.W., Gilbertson, J., Gupta, D., Harrison, J., Legowski, E., Crowley, R.S.: Implementation and evaluation of a negation tagger in a pipeline-based system for information extraction from pathology reports. Medinfo **2004**, 663–667 (2004)
13. Spampinato, C., Kavasidis, I., Aldinucci, M., Pino, C., Giordano, D., Faro, A.: Discovering biological knowledge by integrating high-throughput data and scientific literature on the cloud. Concurrency and Computation: Practice and Experience (2013)
14. Wang, S.J., Middleton, B., Prosser, L.A., Bardon, C.G., Spurr, C.D., Carchidi, P.J., Kittler, A.F., Goldszer, R.C., Fairchild, D.G., Sussman, A.J., et al.: A cost-benefit analysis of electronic medical records in primary care. The American journal of medicine **114**(5), 397–403 (2003)
15. Zhou, X., Han, H., Chankai, I., Prestrud, A., Brooks, A.: Approaches to text mining for clinical medical records. In: Proceedings of the 2006 ACM symposium on Applied computing, pp. 235–239. ACM (2006)
16. Zhou, X., Han, H., Chankai, I., Prestrud, A.A., Brooks, A.D.: Converting semi-structured clinical medical records into information and knowledge. In: 21st International Conference on Data Engineering Workshops, 2005, pp. 1162–1162. IEEE (2005)