Automated Recognition and Post-coordination of Complex Clinical Terms

Philip GOOCH[a,1] and Abdul ROUDSARI[b]
[a] *Centre for Health Informatics, School of Informatics, City University, London, UK*
[b] *School of Health Information Science, University of Victoria, BC, Canada*

Abstract. One of the key tasks in integrating guideline-based decision support systems with the electronic patient record is the mapping of clinical terms contained in both guidelines and patient notes to a common, controlled terminology. However, a vocabulary of pre-coordinated terms cannot cover every possible variation - clinical terms are often highly compositional and complex. We present a rule-based approach for automated recognition and post-coordination of clinical terms using minimal, morpheme-based thesauri, neoclassical combining forms and part-of-speech analysis. The process integrates MetaMap with the open-source GATE framework.

Keywords. natural language processing, interoperability, clinical decision support

Introduction

Application of natural language processing (NLP) techniques for recognition of biomedical and clinical terms has recently been driven by increased demand for information retrieval and extraction tools for standardising and exchanging data between electronic medical record (EMR) systems[1].

While there are a number of approaches to this task, including rule, dictionary, and statistical based approaches[2], there are essentially three steps involved:

1. *recognize* the text string as a possible term (*candidate term* identification)
2. *classify* the candidate term (e.g. chemical compound, part of body, disease)
3. *map* the term to a single concept (*pre-coordination*) or to qualified, multiple concepts (*post-coordination*) within a standardised vocabulary or ontology.

The last step is essential for semantic interoperability of data between EMR systems. For guideline-based clinical decision support (CDS) to provide point of care recommendations within an EMR, patient data must be mapped to the terminology and data model (a *virtual medical record*) employed by the guideline knowledge base[3].

The UMLS Metathesaurus from the National Library of Medicine (NLM)[4] is a large, multi-lingual vocabulary of biomedical concepts classified according to one or more types from a semantic network. The Metathesaurus comprises over 100 reference terminologies, such as HL7 v3, SNOMED CT and LOINC, which have been adopted as international standards for patient data encoding, and form the basis of the information model adopted by at least one formalised guideline model[3].

[1] Corresponding Author. Contact Philip.Gooch.1@city.ac.uk

MetaMap is a tool for discovering UMLS Metathesaurus concepts in free text[5], and is considered to be the 'gold standard' for this task[6]. It allows complex clinical terms to be mapped to individual concepts within UMLS source vocabularies.

Composition of Clinical Terms

The naming of chemical and biological terms frequently involves the use of Latin and Greek *morphemes*. Such terms are known as *neoclassical compounds*. Computational analysis of neoclassical compounds can help identify unknown terms and provide classification for human review[7].

In this paper, we present a purely rule-based application utilising neoclassical combining forms (NCF), part-of-speech (POS) analysis and lexical rules for recognition of complex clinical terms. A post-coordination module annotates the clinical terms with metadata from MetaMap and provides output in HL7 v3 CDA XML.

1. Method

We constructed an NLP pipeline within the GATE framework[8]. The pipeline consists of generic modules for tokenization, POS tagging and noun-phrase identification. Modules were also developed for identification of neoclassical compounds; anatomical terms; chemical nomenclature; proteins, drugs and enzymes; temporal expressions and quantities; and clinical term post-coordination. A MetaMap plugin for GATE was developed in Java using the MetaMap Java API[9]. The transducer modules were written in the GATE JAPE grammar.

1.1. Identification of Potential Clinical Terms

Following approaches suggested by [7],[10], we created lists of neoclassical morphemes and classified them as prefixes, roots and suffixes relating to bodily concepts (e.g. *gastr-*, *haem-*, *derm-*), clinical signs (*cirrh-*, *glauc-*, *-itis*, *-asis*, *-lytic*) and descriptive and positional terms (*ankyl-*, *pachy-*, *inter-*, *intra-*). Suffixes considered to be strongly indicative of a clinical term without an accompanying neoclassical prefix or root were grouped separately (e.g. *-itis*, *-ostomy*).

Regular expressions were written to combine the morphemes into patterns that represent a complete neoclassical compound.

1.2. Recognition of Anatomical Terms

Anatomical terms tend to be highly compositional. We created gazetteers from which to construct regular expressions for identification of complete anatomical structures, using functional prefixes (*adductor*, *extensor*), positional prefixes (*anterior*, *posterior*, *distal*, *dorsal*); anatomical parts and surfaces (*bursa*, *cortex*, *fossa*, *fascia*), organs and organ parts. Complete anatomical terms can then be recognised, for example:

Endothoracic$_{\text{neoclassical}}$ **fascia**$_{\text{part}}$ **of anterior**$_{\text{position}}$ **thoracic**$_{\text{neoclassical}}$ **wall**$_{\text{part}}$

Right$_{\text{position}}$ **posterior**$_{\text{position}}$ **cusp**$_{\text{part}}$ **of aortic**$_{\text{neoclassical}}$ **valve**$_{\text{part}}$

1.3. Recognition of Chemical Compounds

Using lists of chemical element names, ions and alkane prefixes we created combinatorial rules that implement the IUPAC nomenclature[11, 12] for organic and inorganic compounds. Systematic combination of chemical morphemes allows a variety of compounds to be identified, for example:

di$_{\text{multiplier}}$**sodium**$_{\text{element}}$**hypo**$_{\text{oxidation\_prefix}}$**chlor**$_{\text{ion\_root}}$**ite**$_{\text{oxidation\_suffix}}$

tri$_{\text{multiplier}}$**chlor**$_{\text{ion\_root}}$**ometh**$_{\text{alkane\_root}}$**ane**$_{\text{alkane\_suffix}}$

2,3-di$_{\text{multiplier}}$**eth**$_{\text{alkane\_root}}$**yl**$_{\text{alkane\_suffix}}$**pent**$_{\text{alkane\_root}}$**ane**$_{\text{alkane\_suffix}}$

Additional production rules were created by combining chemical morphemes with neoclassical terminals such as *-ase* (enzymes: e.g. acetylcholinesterase), *-ein*, *-in*, *-ine*, *-an* (biological molecules: e.g. ferritin).

1.4. Candidate Term Post-coordination

Prepositional terms comprise two or more noun phrases joined by the prepositions ‘or’ or ‘to’[13], e.g. ‘carcinoma of the lung’.

Such compound phrases require an initial normalisation process so that they can be properly post-coordinated with qualifier concepts from the Metathesaurus.

After experimenting with different noun-phrase combinations to see which combinations gave the best results in MetaMap, we devised the following heuristic:

1. Remove non-negating determiners (*a*, *the*, *this*) from each noun phrase
2. Store the first noun phrase
3. Reverse the order of the remaining noun phrases
4. Add the last token of the first noun phrase to the end
5. Add the remaining tokens of the first noun phrase to the beginning

Example: ‘an ipsilateral fracture of the left femoral neck’

Step 1: {ipsilateral fracture}{left femoral neck}
Step 2: ~~{ipsilateral fracture}~~{left femoral neck}
Step 3: {left femoral neck}
Step 4: {left femoral neck}{fracture}
Step 5: {ipsilateral}{left femoral neck}{fracture}

Although this form is not exactly equivalent semantically to simply reversing the order of the noun phrases[13], this method produces the desired qualified concepts from MetaMap:

```
Meta Mapping (875):
  637 Ipsilateral {SNOMEDCT} [Spatial Concept]
  637 Left {SNOMEDCT} [Spatial Concept]
  884 Femoral Neck Fracture (Femoral neck fracture {SNOMEDCT}) [Injury or
Poisoning]
```

From this HL7v3 CDA XML can be generated by mapping UMLS semantic types to their corresponding qualifiers in SNOMED.

2. Results

The NCF transducer was evaluated against a corpus of 500 MedLine abstracts that had previously been annotated solely with MetaMap (Table 1).

Table 1. Recall and precision of neoclassical combining forms(B) vs MetaMap(A): whole abstract

Annotation	Match	Only A	Only B	Overlap	Recall	Precision	F1.0lenient
Medical_Term	7682	14953	1358	4466	0.45	0.90	0.60

Candidate terms identified by the neoclassical rules were submitted to MetaMap for concept mapping; comparing only the candidate terms against those validated by MetaMap yielded the results shown in Table 2.

Table 2. Recall and precision of neoclassical combining forms(B) vs MetaMap(A): candidate terms only

Annotation	Match	Only A	Only B	Overlap	Recall	Precision	F1.0lenient
Medical_Term	9597	861	1358	2551	0.93	0.90	0.92

The performance of the entire pipeline has not yet been formally evaluated: we are currently refining the rules against an NLP research data set from i2b2[14].

3. Discussion

The high precision but moderate recall of the NCF rules shows that they are useful but insufficient for identifying clinical terms in unstructured text. Additional rules are required for matching anatomical terms, biochemical compounds and complex phrases.

3.1. Related Work

[15] presented a module for recognising medical terms using neoclassical forms, although they validated candidate terms against a general lexicon (EuroWordNet), rather than a specialist biomedical thesaurus. [10] suggested a methodology for medical term recognition using NCFs, but an implementation was not described.

[13] proposed a similar approach to normalising prepositional terms, although they used simple noun-phrase order inversion. However, our approach seemed to produce more useful results from MetaMap, although this requires more formal evaluation.

[16] developed the open-source Health Information Text Extraction (HITEx) tool using GATE. As with our approach, this combined standard GATE modules for POS tagging and noun-phrase chunking, and a regular expression based term identifier. They used the UMLS Metathesaurus directly, although their tool provided similar concept mapping functionality to MetaMap. Additionally, they used a machine-learning component, currently missing from our approach. It is not clear whether they used neoclassical combining forms to assist in term identification. Also, their rules are written in a compact, undocumented syntax, whereas our rules are written in JAPE, which is well documented, easy to modify and extend.

4. Conclusion

We have developed a rules-based approach for clinical term identification, concept mapping and post-coordination within the GATE framework that integrates with MetaMap. Our approach provides high precision with moderate recall when evaluated against general biomedical texts. However, with rule refinement for specific domains (such as clinical notes and clinical guidelines), recall should improve.

We aim to release the code as an open-source project so that the rules can be shared and enhanced by other researchers working in the field.

Acknowledgements

We thank Angus Roberts, Natural Language Processing Group, Department of Computer Science, University of Sheffield, for assistance with GATE and in developing the MetaMap plugin. Deidentified clinical records used in this research were provided by the i2b2 National Center for Biomedical Computing funded by U54LM008748 and were originally prepared for the Shared Tasks for Challenges in NLP for Clinical Data organized by Dr. Ozlem Uzuner, i2b2 and SUNY.

References

[1] W.W. Chapman and K.B. Cohen, *Current issues in biomedical text mining and natural language processing*, *Journal of Biomedical Informatics* **42**(5) (2009), 757-759.

[2] M. Krauthammer and G. Nenadić, *Term identification in the biomedical literature*, *J Biomed Inf* **37** (2004), 512-526.

[3] S.W. Tu, J.R. Campbell, J. Glasgow, M.A. Nyman, et al., *The SAGE Guideline Model: Achievements and Overview*, *Journal of the American Medical Informatics Association* **14**(5) (2007), 589-598.

[4] National Library of Medicine, *UMLS® Reference Manual*, Bethesda, MD, National Library of Medicine, 2009.

[5] A.R. Aronson and F.-M. Lang, *An overview of MetaMap: historical perspective and recent advances*, *J Am Med Inform Assoc* **17** (2010), 229-236.

[6] N.H. Shah, N. Bhatia, C. Jonquet, D. Rubin, et al., *Comparison of concept recognizers for building the Open Biomedical Annotator*, *BMC Bioinformatics* **10(Suppl 9)** (2009), S14.

[7] A.T. McCray, A.C. Browne, and D.L. Moore, *The Semantic Structure of Neo-Classical Compounds*. 1988, National Library of Medicine.

[8] H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan, *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'02)*, Philadelphia, 2002.

[9] National Library of Medicine, *Overview (MetaMap API)*, [Accessed 01 July 2010], Available from: http://mmtx.nlm.nih.gov/javaapi/javadoc/, 2009.

[10] S. Ananiadou, *A methodology for automatic term recognition.*, in *Proceedings of the 15th conference on Computational linguistics*, Kyoto, Japan, Association for Computational Linguistics, 1994.

[11] IUPAC, *Nomenclature of Organic Chemistry, Sections A, B, C, D, E, F, and H*, Oxford, Pergamon Press, 1979.

[12] IUPAC, *A Guide to IUPAC Nomenclature of Organic Compounds (Recommendations 1993)*, Oxford, Blackwell Scientific publications, 1993.

[13] G. Nenadić, S. Ananiadou, and J. McNaught, *Enhancing automatic term recognition through recognition of variation*, in *Proceedings of the 20th international conference on Computational Linguistics*, Geneva, Switzerland, Association for Computational Linguistics Morristown, NJ, USA, 2004.

[14] Ö. Uzuner, Y. Juo, and P. Szolovits, *Evaluating the state-of-the-art in automatic de-identification*, *J Am Med Inform Assoc* **14**(5) (2007), 550-63.

[15] R. Estopà, J. Vivaldi, and M.T. Cabré, *Use of Greek and Latin forms for term detection*, in *Second International Conference on Language Resources and Evaluation*, Athens, Greece, 2000.

[16] Q.T. Zeng, S. Goryachev, S. Weiss, M. Sordo, et al., *Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system BMC Medical Informatics and Decision Making* **6** (2006), 30.