Learning Meronyms from Biomedical Text

Angus Roberts
Department of Computer Science, University of Sheffield,
Regent Court, 211 Portobello Street, Sheffield S1 4DP
a.roberts@dcs.shef.ac.uk

Abstract

The part-whole relation is of special importance in biomedicine: structure and process are organised along partitive axes. Anatomy, for example, is rich in part-whole relations. This paper reports preliminary experiments on part-whole extraction from a corpus of anatomy definitions, using a fully automatic iterative algorithm to learn simple lexico-syntactic patterns from multiword terms. The experiments show that meronyms can be extracted using these patterns. A failure analysis points out factors that could contribute to improvements in both precision and recall, including pattern generalisation, pattern pruning, and term matching. The analysis gives insights into the relationship between domain terminology and lexical relations, and into evaluation strategies for relation learning.

1 Introduction

We are used to seeing words listed alphabetically in dictionaries. In terms of meaning, this ordering has little relevance beyond shared roots. In the OED, `jam` is sandwiched between `jalpaite` (a sulphide) and `jama` (a cotton gown). It is a long way from `bread` and `raspberry`[1]. Vocabularies, however, do have a natural structure: one that we rely on for language understanding. This structure is defined in part by lexical, or sense, relations, such as the familiar relations of synonymy and hyponymy (Cruse, 2000). Meronymy relates the lexical item for a part to that for a whole, equivalent to the conceptual relation of *partOf*[2]. Example 1 shows a meronym. When we read the text, we understand that the `frontal lobes` are not a new entity unrelated to what has gone before, but part of the previously mentioned `brain`.

(1) MRI sections were taken through the brain. Frontal lobe shrinkage suggests a generalised cerebral atrophy.

The research described in this paper considers meronymy, and its extraction from text. It is taking place in the context of the Clinical e-Science Framework (CLEF) project [3], which is developing information extraction (IE) tools to allow querying of medical records. Both IE and querying require domain knowledge, whether encoded explicitly or implicitly. In IE, domain knowledge is required to resolve co-references between textual entities, such as those in Example 1. In querying, domain knowledge is required to expand and constrain user expressions. For example, the query in Example 2 should retrieve sarcomas in the pelvis, but not in limbs.

(2) Retrieve patients on Gemcitabine with advanced sarcomas in the trunk of the body.

The part-whole relation is critical to domain knowledge in biomedicine: the structure and function of biological organisms are organised along partitive axes. The relation is modelled in several medical knowledge resources (Rogers and Rector, 2000),

[1] Oxford English Dictionary, Second Edition, 1989.

[2] Although it is generally held that *partOf* is not just a single simple relation, this will not be considered here.

[3] http://www.clef-user.com/

Proceedings of the ACL Student Research Workshop, pages 49–54,
Ann Arbor, Michigan, June 2005. ©2005 Association for Computational Linguistics

but they are incomplete, costly to maintain, and unsuitable for language engineering. This paper looks at simple lexico-syntactic techniques for learning meronyms. Section 2 considers background and related work; Section 3 introduces an algorithm for relation extraction, and its implementation in the PartEx system; Section 4 considers materials and methods used for experiments with PartEx. The experiments are reported in Section 5, followed by conclusions and suggestions for future work.

2 Related Work

Early work on knowledge extraction from electronic dictionaries used lexico-syntactic patterns to build relational records from definitions. This included some work on *partOf* (Evens, 1988). Lexical relation extraction has, however, concentrated on hyponym extraction. A widely cited method is that of Hearst (1992), who argues that specific lexical relations are expressed in well-known intrasentential lexico-syntactic patterns. Hearst successfully extracted hyponym relations, but had little success with meronymy, finding that meronymic contexts are ambiguous (for example, `cat's paw` and `cat's dinner`). Morin (1999) reported a semi-automatic implementation of Hearst's algorithm. Recent work has applied lexical relation extraction to ontology learning (Maedche and Staab, 2004).

Berland and Charniak (1999) report what they believed to be the first work finding part-whole relations from unlabelled corpora. The method used is similar to that of Hearst, but includes metrics for ranking proposed part-whole relations. They report 55% accuracy for the top 50 ranked relations, using only the two best extraction patterns.

Girju (2003) reports a relation discovery algorithm based on Hearst. Girju contends that the ambiguity of part-whole patterns means that more information is needed to distinguish meronymic from non-meronymic contexts. She developed an algorithm to learn semantic constraints for this differentiation, achieving 83% precision and 98% recall with a small set of manually selected patterns. Others have looked specifically at meronymy in anaphora resolution (e.g. Poesio et al (2002)).

The algorithm presented here learns relations directly between semantically typed multiword terms,

Input:

- A lexicon
- Relations between terms
- Corpus from which to learn

Output:

- New relations
- New terms
- Context patterns

Steps:

1. Using input resources
   (a) Label terms
   (b) Label relations
2. For a fixed number of iterations or until no new relations are learned
   (a) Identify contexts that contain both participants in a relation
   (b) Create patterns describing contexts
   (c) Generalise the patterns
   (d) Use generalised patterns to identify new relation instances
   (e) Label new terms
   (f) Label new relations

Figure 1: PartEx algorithm for relation discovery

and itself contributes to term recognition. Learning is automatic, with neither manual selection of best patterns, nor expert validation of patterns. In these respects, it differs from earlier work. Hearst and others learn relations between either noun phrases or single words, while Morin (1999) discusses how hypernyms learnt between single words can be projected onto multi-word terms. Earlier algorithms include manual selection of initial or "best" patterns. The experiments differ from others in that they are restricted to a well defined domain, anatomy, and use existing domain knowledge resources.

3 Algorithm

Input to the algorithm consists of existing lexical and relational resources, such as terminologies and ontologies. These are used to label text with training relations. The context of these relations are found automatically, and patterns created to describe these contexts. These patterns are generalised and used to discover new relations, which are fed back iteratively into the algorithm. The algorithm is given in Figure 1. An example iteration is shown in Figure 2.

3.1 Discovering New Terms

Step 2e in Figure 1 labels new terms, which may be discovered as a by-product of identifying new rela-

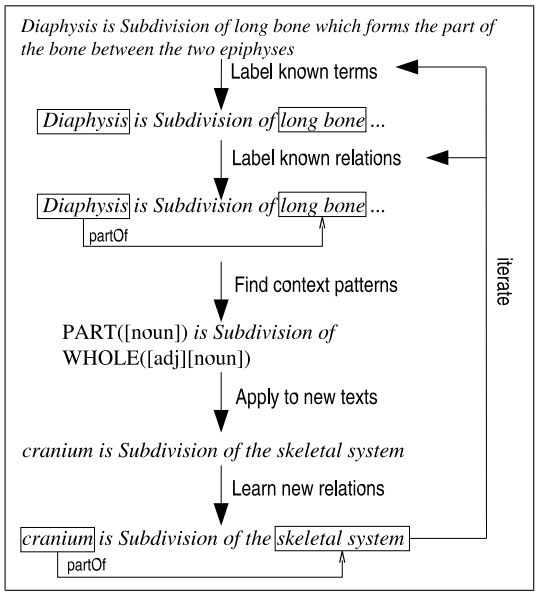

Figure 2: PartEx relation discovery between terms, patterns represented by tokens and parts of speech.

tion instances. This is possible because there is a distinction between the lexical item used to find the pattern context (Step 2a), and the pattern element against which new relations are matched (Step 2d). For example, a pattern could be found from the context (`term` *relation* `term`), and expressed as (`noun` *relation* `adjective noun`). When applied to the text to learn new relation instances, sequences of tokens taking part in this relation will be found, and may be inferred to be terms for the next iteration.

3.2 Implementation: PartEx

Implementation was independent of any specific relation, but configured, as the PartEx system, to discover *partOf*. Relations were usually learned between terms, although this was varied in some experiments. The algorithm was implemented using the GATE NLP framework (Cunningham et al., 2002) and texts preprocessed using the tokeniser, sentence splitter, and part-of-speech (POS) tagger provided with GATE. In training, terms were labelled using MMTx, which uses lexical variant generation to map noun phrases to candidate terms and concepts attested in a terminology database. Final candidate selection is based on linguistic matching metrics, and concept resolution on filtering ambiguity from the MMTx source terminologies (Aronson, 2001). Training relations were labelled from an existing meronymy. Simple contexts of up to five tokens between the participants in the relation were identified using JAPE, a regular expression language integrated into GATE. For some experiments, relations were considered between noun phrases, labelled using LT CHUNK (Mikheev and Finch, 1997). GATE wrappers for MMTx, LT CHUNK, and other PartEx modules are freely available [4].

Patterns describing contexts were expressed as shallow lexico-syntactic patterns in JAPE, and a JAPE transducer used to find new relations. A typical pattern consisted of a sequence of parts of speech and words. Pattern generalisation was minimal, removing only those patterns that were either identical to another pattern, or that had more specific lexico-syntactic elements of another pattern. To simplify pattern creation for the experiments reported here, patterns only used context between the relation participants, and did not use regular expression quantifiers. New terms found during relation discovery were labelled using a finite state machine created with the Termino compiler (Harkema et al., 2004).

4 Materials and Method

Lexical and relational resources were provided by the Unified Medical Language System (UMLS), a collection of medical terminologies [5]. Term lookup in the training phase was carried out using MMTx. Experiments made particular use of The University of Washington Digital Anatomist Foundational Model (UWDA), a knowledge base of anatomy included in UMLS. Relation labelling in the training phase used a meronymy derived by computing the transitive closure of that provided with the UWDA.

The UWDA gives definitions for some terms, as headless phrases that do not include the term being defined. A corpus was constructed from these, for learning and evaluation. This corpus used the first 300 UWDA terms with a definition, that had a UMLS semantic type of “Body Part”. These terms included synonyms and orthographic variants given the same definition. Complete definitions were constructed by prepending terms to definitions with the copula “is”. An example is shown in Figure 2.

[4] http://www.dcs.shef.ac.uk/∼angus

[5] Version 2003AC, http://www.nlm.nih.gov/research/umls/

Experiments were carried out using cross validation over ten random unseen folds, with 71 unique meronyms across all ten folds. Definitions were pre-processed by tokenising, sentence splitting, POS tagging and term labelling. Evaluation was carried out by comparison of relations learned in the held back fold, to those in an artificially generated gold standard (described below). Evaluation was type based, rather than instance based: unique relation instances in the gold standard were compared with unique relation instances found by PartEx, i.e. identical relation instances found within the same fold were treated as a single type. Evaluation therefore measures domain knowledge discovery.

Gold standard relations were generated using the same context window as for Step 2a of the algorithm. Pairs of terms from each context were checked automatically for a relation in UWDA, and this added to the gold standard. This evaluation strategy is not ideal. First, the presence of a part and a whole in a context does not mean that they are being meronymically related (for example, "found in the hand and finger"). The number of spurious meronyms in the gold standard has not yet been ascertained. Second, a true relation in the text may not appear in a limited resource such as the UWDA (although this can be overcome through a failure analysis, as described in Section 4.1). Although a better gold standard would be based on expert mark up of the text, the one used serves to give quick feedback with minimal cost. Standard evaluation metrics were used. The accuracy of initial term and relation labelling were not evaluated, as these are identical in both gold standard creation and in experiments.

4.1 Failure Analysis

For some experiments, a failure analysis was carried out on missing and spurious relations. The reasons for failure were hypothesised by examining the sentence in which the relation occurred, the pattern that led to its discovery, and the source of the pattern.

Some spurious relations appeared to be correct, even though they were not in the gold standard. This is because the gold standard is based on a resource which itself has limits. One of the aims of the work is to supplement such resources: the algorithm *should* find correct relations that are not in the resource. Proper evaluation of these relations requires care, and methodologies are currently being investigated. A quick measure of their contribution was, however, found by applying a simple methodology, based on the source texts being definitional, authoritative, and describing relations in unambiguous language. The methodology adjusts the number of spurious relations, and calculates a *corrected precision*. By leaving the number of actual relations unchanged, corrected precision still reflects the proportion of discovered relations that were correct relative to the gold standard, but also reflects the number of correct relations not in the gold standard. The methodology followed the steps in Figure 3.

1. Examine the context of the relation.
2. If the text gives a clear statement of meronomy, the relation is not spurious.
3. If the text is clearly not a statement of meronomy, the relation is spurious.
4. If the text is ambiguous, refer to a second authoritative resource[6]. If this gives a clear statement of meronomy, the relation is not spurious.
5. If none of these apply, the relation is spurious.
6. Calculate corrected precision from the new number of spurious relations.

Figure 3: Calculating corrected precision.

5 Experimental Results

Table 3 shows the results of running PartEx in various configurations, and evaluating over the same ten folds. The first configuration, labelled BASE, used PartEx as described in Section 3.2, to give a recall of 0.80 and precision of 0.25. A failure analysis for this configuration is given in Table 2. It shows that the largest contribution to spurious relations (i.e. to lack of precision), was due to relations discovered by some pattern that is ambiguous for meronymy (category PATTERN). For example, the pattern "`[noun]` *and* `[noun]`" finds the incorrect meronym "`median` *partOf* `lateral`" from the text "median and lateral glossoepiglottic folds". The algorithm learned the pattern from a correct meronym, and applying it in the next iteration, learned spurious relations, compounding the error.

[6] In this case, Clinically Oriented Anatomy. K. Moore and A. Dalley. 4th Edition. 1999. Lippincott Williams and Wilkins.

Category	Description	Count	%
SPECIFIC	There are one or more variant patterns that come close to matching this relation, but none specific to it.	10	50%
DISCARD	Patterns that could have picked these up were discarded, as they were also generating spurious patterns.	7	35%
SCARCE	The context is unique in the corpus, and so a pattern could not be learnt without generalisation.	3	15%
COMPOUND	The relation is within a compound noun. These are not recognised by the discovery algorithm.	1	5%
COMPLEX	Complex context, which is beyond the simple current "part token* whole" context.	1	5%

Table 1: Failure analysis of 20 missing relations over ten folds, using PartEx configuration FILT.

Category	Description	BASE		FILT	
		Count	%	Count	%
PATTERN	The pattern used to discover the relation does not encode partonomy in this case (Patterns involving: *is* 33 (69%); *and* 10 (21%); *or* 3 (6%); other 2 (4%)).	48	43%	0	0%
CORRECT	Although not in the gold standard, the relation is clearly correct, either from an unambiguous statement of fact in the text from which it was mined, or by reference to a standard anatomy textbook.	30	27%	33	49%
DEEP	The relation is within a deeper structure than the surface patterns considered. The algorithm has found an incorrect relation that relates to this deep structure. For example, the text "limen nasi is subdivision of surface of viscerocranial mucosa" leads to (`limen nasi` *partOf* `surface`).	12	11%	14	21%
FRAGMENT:DEEP	A combination of the FRAGMENT and DEEP categories. For example, given the text "nucleus of nerve is subdivision of neural tree", it has learnt that (`subdivision` *partOf* `neural`).	10	9%	4	6%
FRAGMENT	The relation is a fragment of one in the text. For example, "plica salpingopalatine is subdivision of viscerocranial mucosa" leads to (`plica salpingopalatine` *partOf* `viscerocranial`).	9	8%	12	18%
OTHER	Other reason.	4	4%	3	5%

Table 2: Failure analysis of spurious part-whole relations found by PartEx, for configuration BASE (over half the spurious relations across ten folds) and configuration FILT (all spurious relations in ten folds). In each case, a small number of relations are in two categories.

	Possible	Actual	Missing	Spurious	P	R
BASE	71	56	15	168	0.25	**0.80**
FILT	71	51	20	67	0.43	0.73
CORR	71	51	20	34	**0.58**	0.73
ITR1	71	45	26	66	0.39	0.62
ITR2	71	51	20	67	0.43	0.73
TERM	71	51	20	213	0.20	0.74
TOK	30	26	4	266	0.09	0.88
NP	32	27	5	393	0.07	0.81
POS	71	21	50	749	0.03	0.32

Table 3: Evaluation of PartEx. Total number of relations, mean precision (P) and mean recall (R) for various configurations, as discussed in the text.

The bulk of the spurious results of this type were learnt from patterns using the tokens *and, is*, and *or*.

This problem needs a principled solution, perhaps based on pruning patterns against a held-out portion of training data, or by learning ambiguous patterns from a large general corpus. Such a solution is being developed. In order to mimic it for the purpose of these experiments, a filter was built to remove patterns derived from problematic contexts. Table 3 shows the results of this change, as configuration FILT: precision rose to 0.43, and recall dropped. All other experiments reported used this filter.

A failure analysis of missing relations from configuration FILT is shown in Table 1. The drop in recall is explained by PartEx filtering ambiguous patterns. The biggest contribution to lack of recall was over-specific patterns (for example, the pattern "`[term]` *is part of* `[term]`" would not identify the meronym in "finger is a part of the hand". Generalisation of patterns is essential to improve recall. Improvements could also be made with more sophisticated context, and by examining compounds.

A failure analysis of spurious relations for configuration FILT is shown in Table 2. The biggest impact on precision was made by relations that could be considered correct, as discussed in Section 4.1. A corrected precision of 0.58 was calculated, shown as configuration CORR in Table 3. Two other factors affecting precision can be deduced from Table 2. First, some relations were encoded in deeper linguistic structures than those considered (category DEEP). Improvements could be made to precision by considering these deeper structures. Second, some spurious relations were found between fragments of terms, due to failure of term recognition.

The algorithm used by PartEx is iterative, the implementation completing in two iterations. Configurations ITR1 and ITR2 in Table 3 show that both recall and precision increase as learning progresses.

Four other experiments were run, to assess the impact of term recognition. Results are shown in Table 3. Configuration TERM continued to label terms in the training phase, but did not label new terms found during iteration (as discussed in Section 3.1).

TOK and NP used no term recognition, instead finding relations between tokens and noun phrases respectively (the gold standard being amended to reflect the new task). POS omitted part-of-speech tags from patterns. In all cases, there was a large increase in spurious results, impacting precision. Term recognition seemed to provide a constraint in relation discovery, although the nature of this is unclear.

6 Conclusions

The PartEx system is capable of fully automated learning of meronyms between semantically typed terms, from the experimental corpus. With simulated pattern pruning, it achieves a recall of 0.73 and a precision of 0.58. In contrast to earlier work, these results were achieved without manual labelling of the corpus, and without direct manual selection of high performance patterns. Although the cost of this automation is lower results than the earlier work, failure analyses provide insights into the algorithm and scope for its further improvement.

Current work includes: automated pattern pruning, extending pattern context and generalisation; incorporating deeper analyses of the text, such as semantic labelling (c.f. Girju (2003)) and the use of dependency structures; investigating the rôle of term recognition in relation discovery; measures for evaluating new relation discovery; extraction of putative sub-relations of meronymy. Work to scale the algorithm to larger corpora is also under way, in recognition of the fact that the corpus used was small, highly regularised, and unusually rich in meronyms.

Acknowledgements

This work was supported by a UK Medical Research Council studentship. The author thanks his supervisor Robert Gaizauskas for useful discussions, and the reviewers for their comments.

References

A. Aronson. 2001. Effective Mapping of Biomedical Text to the UMLS Metathesaurus: The MetaMap Program. In *Proceedings of the 2001 American Medical Informatics Association Annual Symposium*, pages 17–21, Bethesda, MD.

M. Berland and E. Charniak. 1999. Finding Parts in Very Large Corpora. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 57–64, College Park, MD.

D. Cruse. 2000. *Meaning in Language: An Introduction to Semantics and Pragmatics*. Oxford University Press.

H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. 2002. 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*, pages 168–175, Philadelphia, PA.

M. Evens, editor. 1988. *Relational Models of the Lexicon: Representing Knowledge in Semantic Networks*. Cambridge University Press.

R. Girju, A. Badulescu, and D. Moldovan. 2003. Learning Semantic Constraints for the Automatic Discovery of Part-Whole Relations. In *Proceedings of the Human Language Technology Conference / North American Chapter of the Association for Computational Linguistics Conference*, Edmonton, Canada.

H. Harkema, R. Gaizauskas, M. Hepple, N. Davis, Y. Guo, A. Roberts, and I. Roberts. 2004. A Large-Scale Resource for Storing and Recognizing Technical Terminology. In *Proceedings of 4th International Conference on Language Resources and Evaluation*, Lisbon, Portugal.

M. Hearst. 1992. Automatic Acquisition of Hyponyms from Large Text Corpora. In *Proceedings of the Fourteenth International Conference on Computational Linguistics*, pages 539–545, Nantes, France.

A. Maedche and S. Staab. 2004. Ontology Learning. In *Handbook on Ontologies*, pages 173–190. Springer.

A. Mikheev and S. Finch. 1997. A Workbench for Finding Structure in Texts. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, pages 372–379, Washington D.C.

E. Morin and C. Jacquemin. 1999. Projecting Corpus-based Semantic Links on a Thesaurus. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 389–396, College Park, MD.

M. Poesio, T. Ishikawa, S. Schulte im Walde, and R. Vieira. 2002. Acquiring Lexical Knowledge for Anaphora Resolution. In *Proceedings of the Third International Conference on Language Resources and Evaluation*, Las Palmas, Canary Islands.

J. Rogers and A. Rector. 2000. GALEN's Model of Parts and Wholes: Experience and Comparisons. In *Proceedings of the 2000 American Medical Informatics Association Annual Symposium*, pages 714–718, Philadelphia, PA.