Unsupervised Relation Extraction for Automatic Generation of Multiple-Choice Questions

Naveed Afzal
Research Institute for Information
and Language Processing
University of Wolverhampton
Wolverhampton, UK
n.afzal@wlv.ac.uk

Viktor Pekar
Oxford University Press
Great Clarendon St.
Oxford, OX2 6DP, UK
viktor.pekar@oup.com

Abstract

In this paper, we investigate an unsupervised approach to Relation Extraction to be applied in the context of automatic generation of multiple-choice questions (MCQs). The approach aims to identify the most important semantic relations in a document without assigning explicit labels to them in order to ensure broad coverage, unrestricted to predefined types of relations. The paper examines three different surface pattern types, each implementing different assumptions about linguistic expression of semantic relations between named entities. Our main findings indicate that the approach is capable of achieving high precision rates and its enhancement with linguistic knowledge helps to produce significantly better patterns. The intended application for the method is an e-learning system for automatic assessment of students' comprehension of training texts; however it can also be applied to other NLP scenarios, where it is necessary to recognise important semantic relations without any prior knowledge as to their types.

Keywords

Information Extraction, Relation Extraction, Biomedical domain, MCQ generation.

1. Introduction

Information Extraction (IE) is an important problem in many information access applications. The goal is to identify instances of specific semantic relations between named entities of interest in the text. As is known from the literature, Relation Extraction in the biomedical domain is quite difficult compared to other domains, such as news domain, due to the inherently complex nature of its texts: biomedical Named Entities (NEs) are expressed in various linguistic forms such as abbreviations, plurals, compounds, coordination, cascades, acronyms and apposition. Sentences in such texts are syntactically complex as the subsequent Relation Extraction phase depends upon the correct identification of the named entities and correct analysis of linguistic constructions expressing relations between them (e.g., [3, 21]).

The main advantage of the approach presented in this paper is that it can cover a potentially unrestricted range of semantic relations while most supervised and semi-supervised approaches can learn to extract only those relations that have been exemplified in annotated text, seed patterns or seed named entities. Moreover, our approach is suitable in situations where a lot of unannotated text is available as it does not require manually annotated text or seeds. These properties of the method can be useful, specifically, in such applications as Multiple-Choice Question generation [12] or a pre-emptive approach in which viable IE patterns are created in advance without human intervention [20,15].

In the future, we plan to employ the Relation Extraction method for automatic MCQ generation, where it will be used to find relations and named entities in educational texts that are important for testing students' familiarity with key facts contained in the texts. In order to achieve this, we need an IE method that has a high precision and at the same time works with unrestricted semantic types of relations (i.e. without reliance on seeds), while recall is of secondary importance to precision.

2. Related Work

There is a large body of research dedicated to the problem of extracting relations from general-domain texts, and from biomedical texts in particular. Most previous work focused on supervised methods and tried to both extract relations and assign labels describing their semantic types [16 and 5, among many others]. As a rule, these approaches required a manually annotated corpus, which is very laborious and time-consuming to produce.

Semi-supervised and unsupervised approaches relied on seeds patterns and/or examples of specific types of relations [1, 17, 20, and 15]. They often employ bootstrapping techniques which use a small set of seeds in order to start the learning process. An unsupervised approach based on clustering of candidate patterns for the discovery of the most important relation types among NEs from a newspaper domain was presented by [6]. In the biomedical domain, most approaches were supervised and relied on regular expressions to learn patterns [4], while semi-supervised approaches exploited pre-defined seed patterns and cue words [2, 7, 11].

Supervised approaches or those based on manually-written extraction rules that have been previously used for Relation Extraction in the biomedical domain are

International Conference RANLP 2009 - Borovets, Bulgaria, pages 1–5

inadequate in scenarios where relation types of interest are not known in advance. In the following section, we describe our method for finding such relations in an unsupervised manner.

3. Extraction of candidate patterns

Our general approach to the discovery of interesting extraction patterns consists of two main stages: (i) the construction of potential patterns from an unannotated domain corpus and (ii) their relevance ranking.

3.1 Pre-processing steps

The first step in constructing candidate patterns is to perform part-of-speech tagging and NE recognition in an unannotated domain corpus. To do that, we employed the Genia[1] tagger. The Genia tagger tags the following five types of biomedical named entities: Protein, DNA, RNA, Cell Type, and Cell Line. The Genia PoS tagger has been reported to achieve over 96% accuracy on a general corpus (Wall Street Journal) and over 98% on the biomedical Genia corpus [18, 19].

3.2 Linguistic types of patterns

Once the training corpus has been tagged with the Genia tagger, the process of pattern building takes place. Its goal is to identify which NEs are likely to be semantically related to each other. The procedure for constructing candidate patterns is based on the idea that important semantic relations are expressed with the help of recurrent linguistic constructions, and these constructions can be recognised by examining sequences of content words (nouns, verbs, adjectives and adverbs) appearing between NEs. To find such constructions, we impose a limit on the number of content words intervening between two NEs. We experimented with different thresholds and finally settled on minimum one content word and maximum three content words to be extracted between two NEs. The reason for introducing this condition is that if there are no content words between two NEs then, although some relation might exist between them, it is likely to be a very abstract grammatical relation. For example, in "X of Y" there is a relation between X and Y, but the phrase does not explicitly express any domain-specific knowledge. On the other hand, if there are too many content words intervening between two NEs, then it is likely they are not related at all. We build patterns using this approach and store each pattern along with its frequency in a database. In this paper we describe experiments with three different pattern types:

1. Untagged word patterns
2. PoS-tagged word patterns
3. Verb-centred patterns

Untagged word patterns consist of named entities and the content words intervening between them. The reason for choosing these different types of surface patterns is that verbs typically express semantic relations between nouns that are used as their arguments. Some examples of untagged word patterns along with their frequencies are shown in Table 1. Table 2 (*PoS-tagged word patterns)* contains the PoS of each content word, while Table 3 (*verb-centred patterns*) contains patterns where the presence of a verb is compulsory in each pattern. We require the presence of a verb in the verb-based patterns as verbs are the main predicative class of words, expressing specific semantic relations between two named entities.

Table 1: Examples of untagged word patterns

Patterns	Frequency
PROTEIN activation PROTEIN	53
DNA contain DNA	46
PROTEIN bind DNA	39
CELL_TYPE express PROTEIN	31

Table 2: Examples of PoS-tagged patterns

Patterns	Frequency
PROTEIN activation_n PROTEIN	53
PROTEIN include_v PROTEIN	43
PROTEIN activate_v PROTEIN	32
DNA encode_v PROTEIN	27

Table 3: Examples of verb-centred patterns

Patterns	Frequency
PROTEIN bind_v DNA	39
PROTEIN induce_v PROTEIN	29
PROTEIN express_v CELL_TYPE	19
PROTEIN stimulate_v CELL_LINE	11

Moreover, in the pattern building phase, patterns containing passive forms of the verb like:

PROTEIN be_v express_v CELL_TYPE

are converted into the active voice form of the verb:

CELL_LINE express_v PROTEIN

Because such patterns were taken to express a similar semantic relation between NEs, passive to active conversion was carried out in order to relieve the problem of data sparseness: it helped to increase the frequency of unique patterns and reduce the total number of patterns. For the same reason, negation expressions (not, does not, etc) were also removed from the patterns as they express a semantic relation between NEs equivalent to one expressed in patterns where a negation particle is absent.

4. Pattern Ranking

After candidate patterns have been constructed, the next step is to rank the patterns based on their significance in the domain corpus. The ranking method we use requires

[1] http://www-tsujii.is.s.u tokyo.ac.jp/GENIA/tagger/

a general corpus that serves as a source of examples of pattern use in domain-independent texts. To extract candidates from the general corpus, we treated every noun as a potential named-entity holder and the candidate construction procedure described above was applied to find potential patterns of the three different types in the general corpus. In order to score candidate patterns for domain-relevance, we measure the strength of association of a pattern with the domain corpus as opposed to the general corpus. The patterns are scored using the following methods for measuring the association between a pattern and the domain corpus: Information Gain (IG), Information Gain Ratio (IGR), Mutual Information (MI), Normalised Mutual Information (NMI)[2], Log-likelihood (LL) and Chi-Square (CHI). These association measures were included in the study as they have different theoretical principles behind them: IG, IGR, MI and NMI are information-theoretic concepts while LL and CHI are statistical tests of association.

Information Gain measures the amount of information obtained about domain specialisation of corpus *c*, given that pattern *p* is found in it.

$$IG(p,c) = \sum_{d \in \{c,c'\}} \sum_{g \in \{p,p'\}} P(g,d) \log \frac{P(g,d)}{P(g)P(d)}$$

where *p* is a candidate pattern, *c* – the domain corpus, *p'* – a pattern other than *p*, *c'* – the general corpus, *P*(*c*) – the probability of *c* in "overall" corpus {*c*,*c'*}, and *P*(*p*) – the probability of *p* in the overall corpus.

Information Gain Ratio aims to overcome one disadvantage of IG consisting of the fact that IG grows not only with the increase of dependence between *p* and *c*, but also with the increase of the entropy of *p*. IGR removes this factor by normalizing IG by the entropy of the patterns in the corpora:

$$IGR(p,c) = \frac{IG(g,c)}{-\sum_{g \in \{p,p'\}} P(g) \log P(g)}$$

Pointwise Mutual Information between corpus *c* and pattern *p* measures how much information the presence of *p* contains about *c*, and vice versa:

$$MI(p,c) = \log \frac{P(p,c)}{P(p)P(c)}$$

Chi-Square and Log-likelihood are statistical tests which work with frequencies and rank-order scales, both calculated from a contingency table with observed and expected frequency of occurrence of a pattern in the domain corpus. **Chi-Square** is calculated as follows.

$$x^2(p,c) = \sum_{d \in \{c,c'\}} \frac{(O_d - E_d)^2}{E_d}$$

where *O* is the observed frequency of *p* in domain and general corpus respectively and *E* is the expected frequency of *p* in two corpora.

Log-likelihood is calculated according to the following formula:

$$LL(p,c) = 2\left(O_1 \log\left(\frac{O_1}{E_1}\right) + O_2 \log\left(\frac{O_2}{E_2}\right)\right)$$

where O_1 and O_2 are observed frequencies of *p* in the domain and general corpus respectively, while E_1 and E_2 are its expected frequency values in the two corpora.

In addition to these six measures, we introduce a **meta-ranking** method that combines the scores produced by several individual association measures, in order to leverage agreement between different association measures and downplay idiosyncrasies of individual ones. Because the association functions range over different values (for example, IGR ranges between 0 and 1, and MI between $+\infty$ and $-\infty$), we first normalise the scores assigned by each method[3]:

$$s_{norm}(p) = \frac{s(p)}{\max_{q \in P}(s(q))}$$

where *s*(*p*) is the non-normalised score for pattern *p*, from the candidate pattern set *P*. The normalised scores are then averaged across different methods and used to produce a meta-ranking of the candidate patterns.

Given the ranking of candidate patterns produced by a scoring method, a certain number of highest-ranking patterns can be selected for evaluation. We studied two different ways of selecting these patterns: (i) one based on setting a threshold on the association score below which the candidate patterns are discarded (henceforth, *score-thresholding method*) and (ii) one that selects a fixed number of top-ranking patterns (henceforth, *rank-thresholding method*). During the evaluation, we experimented with different rank- and score-thresholding values.

5. Evaluation

5.1 Experimental data

We used the Genia Corpus as the domain corpus while British National Corpus (BNC) was used as a general corpus. Genia corpus consists of 2,000 abstracts extracted from the MEDLINE containing 18,477 sentences. In the evaluation phase, Genia Event

[2] Mutual Information has a well-known problem of being biased towards infrequent events. To tackle this problem, we normalised the MI score by a discounting factor, following the formula proposed in [9].

[3] Patterns with negative MI scores are discarded.

Annotation corpus[4] is used [8]. It consists of 9,372 sentences.

5.2 Evaluation method

In order to evaluate the quality of the extracted patterns, we examined their ability to capture pairs of related named entities in the manually annotated evaluation corpus, without recognising the type of semantic relation. Selecting a certain number of best-ranking patterns, we measure precision, recall and F-score. To test the statistical significance of differences in the results of different methods and configurations, we used a paired t-test, having randomly divided the evaluation corpus into 20 subsets of equal size; each subset containing 461 sentences on average.

6. Results

Table 4 shows the results of top-ranked patterns for each approach respectively while Table 5 shows the results of the score-thresholding method for each approach respectively (for space considerations, the tables show only precision scores; "Untagged" stands for "untagged word patterns", "PoS" – for "PoS-tagged word patterns", "VC" – for "verb-centred patterns").

Table 4: Precision results of rank-thresholding method

	IG	IGR	MI	NMI	LL	CHI	Meta
Top 100 Ranked Patterns							
Untagged	.56	.62	.33	.68	.62	.74	.69
PoS	.79	.80	.43	.84	.80	.90	.86
VC	.65	.65	.38	.79	.65	.83	.83
Top 200 Ranked Patterns							
Untagged	.55	.55	.30	.54	.55	.63	.56
PoS	.74	.74	.42	.71	.74	.75	.76
VC	.70	.69	.36	.72	.69	.74	.76
Top 300 Ranked Patterns							
Untagged	.53	.52	.34	.53	.52	.56	.55
PoS	.72	.73	.46	.72	.72	.74	.73
VC	.71	.70	.41	.60	.70	.62	.67
Top 400 Ranked Patterns							
Untagged	.51	.53	.33	.49	.53	.52	.50
PoS	.70	.70	.45	.64	.70	.69	.69
VC	.65	.66	.42	.55	.66	.55	.59
Top 500 Ranked Patterns							
Untagged	.51	.51	.32	.47	.51	.49	.48
PoS	.68	.68	.42	.61	.68	.62	.63
VC	.59	.59	.45	.51	.59	.51	.54

Table 5: Precision results of score-thresholding method

	IG	IGR	MI	NMI	LL	CHI	Meta
Threshold score > .06							
Untagged	.68	.68	.34	.34	.68	.72	.33
PoS	.72	.73	.43	.43	.73	.88	.44
VC	.68	.68	.44	.44	.68	.76	.44
Threshold score > .07							
Untagged	.65	.65	.34	.34	.65	.73	.55
PoS	.74	.74	.43	.43	.74	.87	.44
VC	.70	.71	.44	.44	.71	.89	.44
Threshold score > .08							
Untagged	.62	.62	.34	.34	.62	.78	.55
PoS	.71	.71	.43	.43	.71	.92	.72
VC	.66	.69	.44	.44	.69	.88	.76
Threshold score > .09							
Untagged	.57	.57	.34	.34	.57	.82	.56
PoS	.70	.72	.43	.43	.72	.96	.72
VC	.67	.67	.44	.44	.67	.88	.75
Threshold score > .1							
Untagged	.50	.50	.34	.34	.50	.81	.55
PoS	.70	.70	.43	.43	.70	.95	.74
VC	.65	.66	.44	.44	.65	.95	.75
Threshold score > .2							
Untagged	0	0	.34	.34	0	.86	.82
PoS	.86	.86	.43	.44	.86	1.00	.90
VC	.85	.85	.43	.44	.85	1.00	.87

6.1 Ranking methods

In both tables, the results of the best performing ranking method are shown in bold font.

The CHI-score method performs best for the selected 100 top ranked patterns while the meta-ranking method comes out second best in all three patterns types. The difference between CHI-score and the second-best method (meta-ranking) is significant at $p < 0.05$ level. In Table 5, the CHI-score ranking method outperforms all the other ranking methods for all three patterns types while IG, IGR and LL come out second best for most of the thresholding score values. Here also the difference from the second-best ranking method is significant ($p < 0.05$). IG, IGR and LL ranking methods perform quite similarly to each other and in general, there is no statistically significant difference between them. While literature on the topic suggests that IGR performs better than the IG [14, 10], we found that in general there is no statistically significant difference between IG and IGR, IGR and LL in all three pattern types. In both sets of experiments, obviously due to the aforementioned problem, MI performs quite poorly; the normalised version of MI helps to alleviate this problem. Moreover, there exists a statistically significant difference ($p < 0.01$) between NMI and the other ranking methods in all three pattern types.

The meta-ranking method did not improve on the best individual ranking method as expected. In Table 4, the meta- ranking method comes out second best for 100, 200 and 300 top ranked patterns but then its performance decreases. Similarly for thresholding score values it comes out second best for all thresholds greater than 0.09. Moreover, we found that there is a statistically significant difference ($p < 0.05$) between the meta-ranking method and all the other ranking methods for all three patterns types.

6.2 Score vs. rank thresholding

We also find out that score-thresholding method produces better results than rank-thresholding as we are

[4] http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/home/wiki.cgi?page=Event+Annotation

able to achieve up to 100% precision with the former technique.

6.3 Types of patterns

PoS-tagged word patterns and verb-centred patterns perform better than untagged word patterns. Verb-centred patterns work well, because verbs are known to express semantic relations between named entities using syntactic arguments to the verb; PoS-tagged word patterns add important semantic information into the pattern and possibly disambiguate words appearing in the pattern. In order to find out that whether the differences between the three patterns types are statistically significant, we carried out a paired t-test again. We found that there is no statistically significant difference between PoS-tagged word patterns and verb-centred patterns. Apart from IG, IGR and LL there is a statistically significant difference between all the ranking methods of untagged word patterns and PoS-tagged word patterns, untagged word patterns and verb-centred patterns respectively.

6.4 Precision vs. F-measure optimisation

The score-thresholding method achieves higher precision than the rank-thresholding method. High precision is quite important in applications such as MCQ generation. In thresholding scores, it is possible to optimise for high precision (up to 100%), though F-measure is generally quite low. MCQ applications rely on the production of good questions rather than the production of all possible questions, so high precision plays a vital role in such applications.

7. Conclusion

In this paper, we have presented an unsupervised approach for Relation Extraction from surface-based patterns intended to be deployed in an e-Learning system for automatic generation of multiple choice questions. We experimented with three different surface-based approaches and showed that PoS-based and verb-centred patterns achieve higher precision compared to untagged word patterns. We explored different ranking methods and found that the Chi-Square ranking method obtained higher precision than the other ranking methods. We employed two techniques: the rank-thresholding method and score-thresholding method and found that thresholding scores perform better.

For future work, we are going to investigate other meta-ranking methods and carry out a task-embedded evaluation, in the context of the multiple-choice question generation problem.

8. References

[1] Agichtein E. and Gravano L. Snowball: Extracting Relations from Large Plaintext Collections. In Proc. of the 5th ACM International Conference on Digital Libraries (DL-00), 2000.

[2] Blaschke A., Andrade M. A., Ouzounis C., and Valencia A. Automatic Extraction of Biological Information from Scientific Text: Protein-Protein interactions, in Proc. of ISMB99, pp. 60-67, 1999.

[3] Cohen A. M., and Hersh W. R. A Survey of Current Work in Biomedical Text Mining. Briefings in Bioinformatics, pp. 57-71, 2005.

[4] Corney D. P., Jones D., Buxton B., and Langdon W. BioRAT: Extracting Biological Information from Full-length Papers. Bioinformatics, pp. 3206-3213, 2004.

[5] Cunningham H., Maynard D., Bontcheva K., and Tablan V. GATE: A framework and graphical development environment for robust NLP tools and applications. In Proc. of ACL'02, Philadelphia, 2002.

[6] Hasegawa T., Sekine S., and Grishman R. Discovering relations among named entities from large corpora. In Proc. of ACL'04, 2004.

[7] Huang M., Zhu X., Payan G. D., Qu K., and Li M. Discovering patterns to extract protein-protein interactions from full biomedical texts. Bioinformatics, pp. 3604-3612, 2004.

[8] Kim J-D., Ohta T., and Tsujii J. Corpus Annotation for Mining Biomedical Events from Literature, BMC Bioinformatics, 2008.

[9] Lin D. and Pantel P. Concept Discovery from Text. In Proc. of Conference on CL 2002. pp. 577-583. Taipei, Taiwan, 2002.

[10] Manning C. and Schütze H. Foundations of Statistical Natural Language Processing. The MIT Press, Cambridge, US, 1999.

[11] Martin E. P., Bremer E., Guerin G., DeSesa M-C., Jouve O. Analysis of Protein/Protein Interactions through Biomedical Literature: Text Mining of Abstracts vs. Text Mining of Full Text Articles. Berlin: Springer-Verlag, pp. 96-108, 2004.

[12] Mitkov, R., Ha, L. A. and Karamanis, N. A computer-aided environment for generating multiple-choice test items. Natural Language Engineering 12(2). Cambridge University Press, pp. 177-194, 2006.

[13] Ono T., Hishigaki H., Tanigami A. and Takagi T. Automated Extraction of Information on Protein–Protein Interactions from the Biological Literature. Bioinformatics, pp. 155-161, 2001.

[14] Quinlan J.R. Induction of decision trees. Machine Learning, 1(1), pp. 81-106, 1986.

[15] Satoshi Sekine. On-Demand Information Extraction. Proc. of the COLING/ACL. Sydney, pp. 731-738, 2006.

[16] Soderland S. Learning Information Extraction Rules for Semi-Structured and Free Text. Machine Learning, 34(1-3):233–272, 1999.

[17] Stevenson M. and Greenwood. A Semantic Approach to IE Pattern Induction. In Proc. of ACL'05, pages 379-386, 2005.

[18] Tsuruoka Y., Tateishi Y., Kim J-D., Ohta T., McNaught J., Ananiadou S., and Tsujii J. Developing a Robust PoS Tagger for Biomedical Text. Advances in Informatics - 10th Panhellenic Conference on Informatics, LNCS 3746, pp. 382-392, 2005.

[19] Tsuruoka Y. and Tsujii J. Bidirectional Inference with the Easiest-First Strategy for Tagging Sequence Data. Proc. of HLT/EMNLP, pp. 467-474, 2005.

[20] Yusuke Shinyama and Satoshi Sekine. Preemptive Information Extraction using Unrestricted Relation Discovery. Proc. of the HLT Conference of the North American Chapter of the ACL. New York, pp. 304-311, 2006.

[21] Zhou G., Su J., Shen D. and Tan C. Recognizing Name in Biomedical Texts: A Machine Learning Approach. Bioinformatics, pp. 1178-1190, 2004.