Journal of Science and Technology ***49****, Number 5 (2011) 35-46*

A SONG – AUTHOR RELATIONSHIP EXTRACTOR FOR VIETNAMESE NEWS ARTICLES

Hoai-Duc Tuan-Nguyen, Thanh-Quyen Doan-Thi, Mi-Ni Tran-Thi, Bao-Quoc Ho

Department of Information System, Faculty of Information Technology, University of Science Ho Chi Minh City – Vietnam

Received September 30, 2011

ABSTRACT

Relationship extraction is a vital task which is required for many research fields such as Ontology construction or Q – A systems. Good results have been reached for widely used languages such as English while the achievement for Vietnamese is still limited. This paper presents a relationship extractor for Vietnamese which is trained with semi-supervised learning. A set of metrics is employed to qualify extraction result. And a simple NER system is integrated to boost the extractor's efficiency.

Keywords. mesh-free methods; stabilized displacement and equilibrium models; smoothed technique; penalty method.

1. INTRODUCTION

Relationship extraction is one of the most important tasks of Natural Language Processing. In this paper, Relationship is restricted to the semantic relation among two or more concepts, and those concepts are presented in text as words or phrases. A Relationship Extractor, as defined by [5], is a system that takes a non-structured text collection as input and reveals semantic relations as well as all related concepts from the collection as output. Semantic relations are useful for many further processes, such as Ontology construction [10, 11], Q – A system tasks [14, 15], Text-to-Scene conversion [6], and recognizing genes as disease causes [2].

Relationship extraction has received much research concern in many prestigious conferences such as ACM [9, 12], Coling/ACL [10], ... It also takes parts in important international Data-mining projects such as ACE (Automatic Content Extraction), DARPA EELD (Evidence Extraction and Link Discovery), ARDA – QUAINT (Question Answering for Intelligence), ARDA – NIMD (Novel Intelligence from Massive Data), ...

Many research works were conducted for widely-used languages such as English, France... Encouraging results are also achieved in less common languages such as Malay [16], Dutch [14]. However, relationship extraction is still a challenge in Vietnamese since language resources are quite limited. This paper presents a solution for extracting Song – Author relationship from Vietnamese news article. This type of relationship is useful for management of copyright infringement.

2. RELATIONSHIP CLASSIFICATION

As stated in [13], semantic relations are quite diversiform. The type of a semantic relation depends on domain as well as context where the relation is found. According to [17], semantic relations can have the form of either binary relation between 2 concepts (this form can be either named relation or anonymous relation) or multiary relation among 3 or more concepts. We present in this paper 3 common classification systems : WordNet, Roxana Girju and UMLS.

2.1. WordNet

WordNet is an online dictionary for English containing 100,000 concepts of different word types. WordNet classifies the relationships among these concepts into 15 categories: Hypernym, Hyponym, Is-part-of, Has-part, Is-member-of, Has-member, Is-stuff-of, Has-stuff, Cause-to, Entail, Attribute, Synonym, Antonomy, Similarity, Se-also. Detailed information about these categories can be found in [19].

2.2 Roxana Girju

Roxana Girju divides semantic relations into 22 categories including : Is-a, Part-Whole, Cause, Instrument, Make/Produce, Kinship, Possession, Source/From, Purpose, Location/Space, Temporal, Experiencer, Means, Manner, Topic, Beneficiary, Property, Theme, Agent, Depiction, Type, Measure. Among these categories, Is-a and Part-Whole are marked as the most common relationships. Detailed information about these categories can be found in [5].

2.3. Unified Medical Language System - ULMS

ULMS was constructed by National Library Medical of America in 1986 and is enriched year after year. By 2006, ULMS contained 139 lexicons of 17 different languages, with about 1.3 billion concepts. These concepts are divided into 135 semantic types of 2 main branches : Entity and Event. These semantic types are linked by 54 semantic relations of 2 main branches : Hierarchical and Non-hierarchical. Detailed information about these semantic relations can be found in [4].

3. RELATED WORKS

Most relationship extractors use machine learning as their solution. According to [1], features used in relationship extraction can be divided into syntactic features and context features. Supervised learning employs mostly syntactic features while semi-supervised learning employs mostly context features.

3.1. Relationship extraction with supervised machine learning

This approach requires a large enough training data set to train the extractor. This data set must be manually tagged. More precisely, semantic relations in the data set must be properly marked by domain specialists. This approach yields good precision (thanks to specialists' knowledge), but it is also costly and domain dependent.

3.1.1. AutoSlog

AutoSlog [7] uses grammatical information to extract relationship. Firstly, noun phrases and verb phrases are recognized. This can be done thanks to a training text collection with all noun phrases and verb phrases manually tagged. Secondly, each phrases are attached with further grammatical information about their function in the sentences where they are found. E.g. a noun phrase may be a subject of a sentence, a verb phrase may be passive predicative or active predicative, and another noun phrase may be the object of that predicative. A predefined rule set is employed to extract semantic relations between phrases with different grammatical functions. Most relationships are found between subjects and objects via the related predicatives. Detailed information about AutoSlog can be found in [7].

3.1.2. AutoSlog – TS

AutoSlog – TS [8] is an amelioration of AutoSlog. More specifically, the base system is boosted with a set of heuristics which can further process semantic clues around each noun phrase more efficiently. Besides, the learnt rules in AutoSlog are evaluated, the rules with low confidence are discarded. More information about this system is presented in [8].

3.1.3. Dependant Tree

[14] introduces a relationship extractor used for a medical Q – A system of Dutch. The extractor uses ULMS as the relationship classification system. This system focuses on 7 relation types mentioned in ULMS : causes, has_definition, occurs, treats, has_symptom, prevents and diagnoses. The extractor is trained with a data set manually tagged with these 7 relation types to learn relation templates.

Each sentence in raw text is parsed into a structure called dependant tree. The root node represents the sentence. Intermediates nodes represent phrases, accompanied with their functions in the sentence. Leaf nodes represent words in each phrase, accompanied with their POS tags. [14] makes use of ULMS to improve the precision of medical phrase recognition, as well as to

assign a medical semantic type to each phrase. Tuples of the form <Subject, Object> is then extracted from the dependant trees as candidates for the desired relationships. Those Subjects and Objects are generalized to their semantic types in ULMS. The scores of such tuples are statistically computed to eliminate bad tuples. Detailed algorithm is in [14].

3.2. Relationship extraction with semi-supervised machine learning

This approach does not require a very large training data set that is completely manually tagged. Instead, a small set of "seeds" is supplied to boot the learning process. Context is remarkably explored to learn relationship templates from documents. This approach can easily adapt to new domains. However, its precision is often lower than supervised learning.

3.2.1. Dipre (Dual Iterative Pattern Relation Extraction)

Dipre [18] makes use of a set of initial limited seed relations, which is created manually. The system will scan through English documents to find occurrences of these seed relations. The contexts around these occurrences are memorized as relation templates. They are used to recognized new relations.

[18] suggests that a template of a binary relation between concept A and concept B should include the context right before A, right after B and between A and B. This system focuses only on book-author relationship. From just 5 initial seed relations, it automatically generates 15000 new relations. However, this solution is suitable only when a large data set is available for mining. The uncontrolled generated templates can lead to numerous wrong relations.

3.2.2. Snowball

Addressing the drawback of Dipre, Snowball [9] suggests employing Named Entity Recognition – NER to boost the system. This arises from the observation that mistakes are often caused by templates where A (or B) is not actually a concept. NER can help assuring template generation with proper enclosed concepts. Furthermore, Snowball computes confidence for each generated template. Templates with confidences below a predefined threshold will be discarded. This solution depends much on the precision of the NER system. Detailed information about this system can be found in [9].

4. A VIETNAMESE RELATIONSHIP EXTRACTOR

Our system focuses on the relationship between Vietnamese songs and their authors. This type of relationship is helpful to copyright and authorship management. Our approach is similar to [18] in that we use semi-supervised learning to extract binary semantic relations. However, to overcome some drawbacks stated in 3.2.1, we suggest some additional techniques to improve the quality of extracted relations.

4.1. Methodology

Figure 1 illustrates the relationship extraction process. This process starts with a limited set of seed relations. We aim at finding occurrences of these seed relations (in the large training data set) to form relationship templates. These templates is stored and then used to find new relations. All qualified new relations are added to the initial seed set and the process recurs with the enlarged seed set.

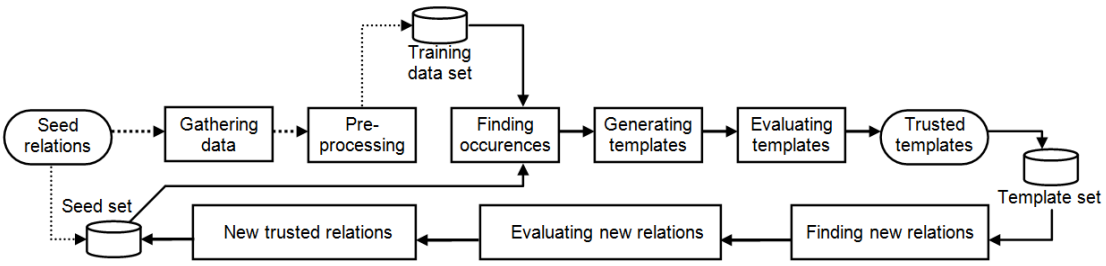

Figure 1. Relationship extraction process

Our approach does not require the training data set to be manually tagged, but it does require this data set to be large enough for the system to do mining effectively. Such large collection is not currently available for Vietnamese. Therefore we perform a phase called Data gathering, i.e. each seed relation is passed into Yahoo Search API as query to find relevant documents. The 100 highest ranked documents for each query are kept as training data. With 20 initial seed relations, we acquire a training data set of 2000 document. Besides, we make use of 1000 documents selected from a Vietnamese news article collection [20]. Hence the final data set includes 3000 documents and is large enough for mining.

Training data is preprocessed to do word segmentation as well as POS tagging (we built our own tools for these tasks since tools for Vietnamese is scarce). Then our system scans through the preprocessed training data set to find occurrences of seed relations. Once an occurrence is found, its context is extracted as relation template.

Each relation template has 6 parts :

- The two related concepts : Song and Author
- Order indicates whether Song or Author is the preceding concept in the sentence
- Prefix is one word right before the preceding concept
- Suffix is one word right after the succeeding concept
- Middle is the string between the 2 concepts

The generated templates are then used to find new semantic relations in the training data set. However, bad templates can give wrong relations. Through the recurrence of this process, if uncontrolled, the number of bad templates as well as wrong relations will terribly increase.

Unlike [18], our system has 2 additional phases called *Templates evaluating* and *New relation evaluating*. These 2 phases is remarkably useful in controlling the quality of templates as well as extracted relations.

4.2. Template Evaluating

Information extraction systems often use different metrics to evaluate learnt templates. [3] suggests 3 metrics of Named entity extraction that can be used for Relationship extraction, including :

- Frequency – fre(P): The times a template P appear in the training dataset. A template which appears many times in the whole training data set is believed to be a reliable template.

- Diversity – div(P): The ratio of distinct relations extracted by template P to all the relations extracted by template P. A template which can generate diversiform relations is believed to be a productive template.

- Specificity – spec(P): The ratio of reasonable relations extracted by template P to all the relations extracted by template P. A Reasonable relation is the relation with the two concepts being actual Song and Author. In order to compute Specificity, a dictionary of Vietnamese songs and authors is required.

In addition to the 3 mentioned metrics, we also make use of the metric suggested by [9] to evaluate relation templates, which is called Template Confident – conf(P). Templates with low confidence will be eliminated. According to [9], relations extracted by a template can be classified into 3 categories: Positive (Both concepts are marked as Song and Author by the NER system), Negative (only one concept is marked as Song / Author by the NER system), Unknown (none of the 2 concepts can be recognized by the NER system). Conf(P) is calculated on the number of Positive and Negative relations extracted by P :

$$Conf(P) = P.Positive / (P.Positive + P.Negative) \quad (1)$$

4.3. New relation Evaluating

New relations extracted after each recurrence are evaluated by their confidences. The confidence of a relation T is computed basing on the confidences of all templates P_0, P_1, P_2, ..., P_n that generate it.

$$Conf(T) = 1 - \prod^{|P|}{}_{i=0}(1 - Conf(P_i)) \quad (2)$$

Relations with confidences above a predefined threshold are kept for the next recurrence. The recurring process stops when no new qualified relation can be found.

4.4. Song – Author dictionary generation

There are 2 reasons leading to the need of a Vietnamese Song – Author dictionary : (i) The template evaluation *Specificity* metric can not be computed without knowing whether concepts

in a relation are actual song / author or not. (ii) Such a dictionary is required by a NER system which can prevent the extractor from generating relations between meaningless "concepts" (i.e. strings being misrecognized as concepts). [9] has proved that integrating a NER system can boost the relationship extractor remarkably. And dictionary lookup is among the fundamental techniques of a basic NER system.

Vietnamese Song – Author dictionary is not currently available. Therefore, we apply the solution suggested by [3] to generate the dictionary. This solution arises from the observation that <list> or <table> tags of an *html* document often enumerate entities of the same types. E.g. if the first 3 rows of a 100-row table present 3 pairs of Song – Author, it is extremely likely that the rest 97 rows also present other Song – Author pairs. Our system explores the World – Wide – Web to look for occurrences of seed Song – Author pairs which reside in the <list> or <table> tags. If at least 3 rows of a <list> or <table> tag contain seed pairs, all other rows are collected to enlarge our seed set. Song – Author pairs in the enlarged seed set are then used to explore the World – Wide – Web in the next recurrence. This process repeats until none of 10 continuous recurrences can yield new Song – Author pairs (i.e. pairs that are not in the dictionary) with the ratio above a threshold. This threshold is experimentally selected to be 5%.

Following the above mentioned process, we generate a dictionary of 14,512 pairs. After manually removing about 10% bad pairs (e.g. pairs with author names such as "unknown" or "being updated"), we have a dictionary of 13,000 qualified Vietnamese Song – Author pairs. This dictionary is used to (i) compute the *Specificity* metric and (ii) support a simple NER system which recognizes entities of type *Song* and type *Author*.

5. EXPERIMENTAL RESULTS

5.1. The data set

The Data gathering phase acquires 2000 news articles from the Internet. This collection joins 1000 news article publish by [20] to form a data set of 3000 documents. We select 300 documents to use as judgment data. The rest of this data set will serve as training data.

No suitable judgment data set is currently available for Vietnamese. So we must construct such data set by manually tag the 300 selected documents. Although this is costly and time consuming, we have no other choice.

The initial seed relation contain 20 Song – Author pairs, Table 1 presents these pairs.

Table 1. List of 20 Vietnamese Song – Author seed relations

Song	**Author**	**Song**	**Author**
Cánh cung	Đỗ Bảo	Hạnh phúc mong manh	Vũ Quốc Việt
Chiếc áo cho em	Lưu Thiên Hương	Đêm nghe tiếng mưa	Đức Trí
Về quê	Phó Đức Phương	Không thể và có thể	Phó Đức Phương
Xuân bên em	Lương Ngọc Châu	Trên đỉnh Phù Vân	Thuận Yến

Song	Author	Song	Author
Người đi xây hồ kẻ gỗ	Nguyễn Văn Tý	Đi qua bóng tối	Minh Tiến
Cỏ hồng	Phạm Duy	Lời chưa nói	Trịnh Thăng Bình
Chơi vơi	Trịnh Công Sơn	Mặt trời bé con	Trần Tiến
Cung đàn xưa	Văn Cao	Nhớ anh	Mỹ Tâm
Một mình	Thanh Tùng	Giấc phù du	Hà Dũng
Vệt nắng cuối trời	Tiến Minh	Tình hoàng hôn	Nguyễn Nhất Huy

5.2. Testing plans and results

After trained, our relationship extractor reaps 2 useful resources : (i) the trusted relation template set and (ii) the trusted seed relation set, including the Vietnamese Song – Author dictionary. Both resources can be used independently or co-operatively to extract new Song – Author relationships from raw text. So we conduct 3 testing plan : Using only the template set, using only the seed set and using both resources. In each plan, we evaluate the efficiency of each metric (mentioned in 4.2 and 4.3) separately.

5.2.1. Plan 1 – Using only the trusted relation template set

Table 2. Experimental result when no metric is used (plan 1)

Precision	54.3%
Recall	62.1%

Table 3. Experimental result when Frequency is used (plan 1)

Threshold	2	3
Precision	55.6%	56.2%
Recall	51.6%	50.3%

Table 4. Experimental result when Diversity is used (plan 1)

Threshold	20%	40%	60%	80%
Precision	53.0%	65.2%	54.5%	8%
Recall	52.3%	39.2%	7.8%	2.6%

Table 5. Experimental result when Specificity is used (plan 1)

Threshold	20%	40%	60%	80%
Precision	59.1%	60.3%	63.3%	65.9%
Recall	61.4%	61.4%	57.5%	55.6%

Table 6. Experimental result when Conf(P) and Conf(T) are used (plan 1)

Conf(P) threshold	20%	20%	20%	20%	40%	40%	40%	40%
Conf(T) threshold	20%	40%	60%	80%	20%	40%	60%	80%
Precision	56.5%	61.1%	60.7%	68.9%	65.6%	65.6%	64.4%	68.8%
Recall	54.2%	50.3%	48.4%	27.5%	38.6%	38.6%	36.6%	21.6%

5.2.2 Plan 2 – Using only the trusted seed relation set

Table 7. Experimental result when no metric is used (plan 2)

Precision	70.1%
Recall	81.0%

Table 8. Experimental result when Frequency is used (plan 2)

Threshold	2	3
Precision	69.9%	72.8%
Recall	80.0%	80.4%

Table 9. Experimental result when Diversity is used (plan 2)

Threshold	20%	40%	60%	80%
Precision	70.5%	68.7%	68.7%	70.1%
Recall	81%	67.3%	67.3%	67.3%

Table 10. Experimental result when Specificity is used (plan 2)

Threshold	20%	40%	60%	80%
Precision	70.7%	71.5%	71.9%	72.8%
Recall	80.4%	80.4%	80.4%	80.4%

Table 11. Experimental result when Conf(P) and Conf(T) are used (plan 2)

Conf(P) threshold	20%	20%	20%	20%	40%	40%	40%	40%
Conf(T) threshold	20%	40%	60%	80%	20%	40%	60%	80%
Precision	74.6%	75.5%	74.3%	90.6%	72.0%	72.0%	71.4%	73.4%
Recall	55.6%	54.2%	51.0%	19%	43.8%	43.8%	42.5%	37.9%

5.2.3. Using both resources

Table 12. Experimental result when no metric is used (plan 3)

Precision	58.8%
Recall	91.5%

Table 13. Experimental result when Frequency is used (plan 3)

Threshold	2	3
Precision	62.6%	73.5%
Recall	89.8%	90.8%

Table 14. Experimental result when Diversity is used (plan 3)

Threshold	20%	40%	60%	80%
Precision	60.9%	67.1%	67.1%	71.1%
Recall	91.5%	70.6%	70.6%	70.6%

Table 15. Experimental result when Specificity is used (plan 3)

Threshold	20%	40%	60%	80%
Precision	62.9%	64.4%	66.5%	69.3%
Recall	90.8%	90.8%	90.8%	90.2%

Table 16. Experimental result when Conf(P) and Conf(T) are used (plan 3)

Conf(P) threshold	20%	20%	20%	20%	40%	40%	40%	40%
Conf(T) threshold	20%	40%	60%	80%	20%	40%	60%	80%
Precision	59.5%	65.6%	66%	73.5%	67.3%	67.3%	66.4%	68.5%
Recall	65.4%	64.7%	64.7%	32.7%	48.4%	48.4%	46.4%	41.2%

5.3. Discussion

Without any metrics employed, recall is quite high but precision is very low. When evaluation metrics are used, precision increases remarkably while recall decrease slightly. The extractor achieves best result with plan 3. The highest recall (91.5%) is acquired when Diversity is used (threshold = 20%), and the highest precision (73.5%) is acquired when Frequency is used (threshold = 3).

6. CONCLUSION AND FUTURE RESEARCH

We develop a Relationship extractor which uses semi-supervised learning. We also employ different metrics to evaluate relation templates as well as extracted relations in order to improve extraction precision. Our system gets encouraging result and does its best with the use of both trusted relation template set and trusted seed relation set. The NER system used in our system is very simple. In future we aim at developing a more efficient NER system to boost our extractor.

REFERENCES

1. Abdulrahman Almuhareb, Massimo Poesio – "MSDA: Wordsense Discrimination Using Context Vectors and Attributes". University of Essex, 2006.
2. Barbara Rosario - "Extraction of semantic relations from bioscience text". University of Trieste, Italy, 1995.
3. Casey Whitelaw, Alex Kehlenbeck, Nemanja Petrovic, Lyle Ungar - Web-Scale Named Entity Recognition, University Pennsylvania, 2008.
4. Bao-Ha Chau-Thi – Mô hình lập chỉ mục trên khái niệm (Ứng dụng cho tìm kiếm thông tin Y học) – Master Thesis. University of Science, HCM City, Vietnam, 2007.
5. Corina Roxana Girju - Text mining for semantic relations. PhD. Thesis. The University of Texas at Dallas, 2002.
6. Coyle B., and Sproat R. – WordsEye: An automatic text-to-scene conversion system, The Siggraph Conference, Los Angeles, USA, 2001.
7. Ellen Riloff – Automatically constructing a dictionary for information extraction tasks, In proceeding of the eleventh national conference on artificial intelligence, 1993, pp. 811–816.
8. Ellen Riloff – Automatically generating extraction patterns from untagged text, In proceeding of the eleventh national conference on artificial intelligence, 1993, pp. 811–816.
9. Eugene Agichtein and Luis Gravano - Snowball: Extracting relations from large plaintext collections, In Proceedings of the Fifth ACM International Conference on Digital Libraries, 2000.
10. Fabian M. Suchanek, Georgiana Ifrim, Gerhard Weikum - LEILA: Learning to Extract Information by Linguistic Analysis, COLING/ACL 2006. (Workshop On Ontology Learning And Population), 2006.
11. Fabian M. Suchanek, Gjergji Kasneci, Gerhard Weikum - YAGO: A Large Ontology from Wikipedia and WordNet, Web Semantics: Science, Services and Agents on the World Wide Web **6** (3) (2008) 203-217.
12. Giuliano C., Lavelli A., Romano L. - Relation extraction and the influence of automatic

named-entity recognition, ACM Transactions on Speech and Language Processing (TSLP) **5** (1) (2007).

13. Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid O. Seaghdha, Sebastian Pado, Marco Pennacchiotti, Lorenza Romano and Stan Szpakowicz - Multi-Way Classification of Semantic Relations Between Pairs of Nominals, The NAACL-HLT-09 Workshop on Semantic Evaluations: Recent Achievements and Future Directions (SEW-09), Boulder, USA, May 2009.
14. Ismail Fahmi - Automatic Term and Relation Extraction for Medical Question Answering System, Center for Language and Cognition Groningen (CLCG) of the Faculty of Arts of the University, 2009.
15. Kim S., Lewis P., Martinez K. and Goodall S. – Question Answering Towards Automatic Augmentations of Ontology Instances, The Semantic Web: Research and Applications: First European Semantic Web Symposium, ESWS: 152-166, 2004.
16. Mohd Juzaiddin Ab Aziz, Fatimah Dato'Ahmad, Abdul Azim Abdul, Ramlan Mahmod Ghani – Pola grammar technique for grammatical relation extraction in malay language, Fakulti Sains Komputer & Teknologi Maklumat, Universiti Putra Malaysia 43400 Serdang,Selangor, Malaysia and Fakulti Teknologi & Sains Maklumat, Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor, Malaysia, 2009.
17. Nguyen Bach, Sameer Badaskar – "A survey on relation extraction (2008)" – Language Tchnology Institute, University of Carnegie Mallon.
18. Sergey Brin – Extracting patterns and relations from the world wide web, In WebDB Workshop at 6th International Conference on Extending Database Technology, EDBT'98, 1998.
19. http://wordnet.princeton.edu/
20. http://www.fit.hcmus.edu.vn/~ddien/VnTCCor/

Corresponding author:

Hoai-Duc Tuan-Nguyen,

Department of Information System,

Faculty of Information Technology, University of Science

Email: *tnhduc@fit.hcmus.edu.vn*