Automatic Identification of Concepts and Conceptual relations from Patents Using Machine Learning Methods

Pattabhi RK Rao
AU-KBC Research Centre
MIT, Anna University
Chrompet, Chennai, India

Sobha Lalitha Devi
AU-KBC Research Centre
MIT, Anna University
Chrompet, Chennai, India
sobha@au-kbc.org

Paolo Rosso
Natural Language Engineering Lab
Technical University of Valencia
Spain
prosso@dsic.upv.es

Abstract

This paper presents a machine learning approach to automatically extract concepts and the conceptual relations towards creation of Conceptual Graphs (CGs) from patent documents using shallow parser and NER. The main challenge in the creation of conceptual graphs from the natural language texts is the automatic identification of concepts and conceptual relations. The texts analyzed in this work are patent documents, focused mainly on the claim's section (Claim) of the documents. The task of automatically identifying the concept and conceptual relation becomes difficult due to the complexities in the writing style of these documents as they are technical as well as legal. The analysis we have done shows that the general in-depth parsers available in the open domain fail to parse the 'claims section' sentences in patent documents. The failure of in-depth parsers led us, to develop a methodology to extract CGs using other resources. Thus in the present work we came up with a methodology which uses shallow parsing, NER and machine learning technique for extracting concepts and conceptual relationships from sentences in the claim/novelty section of patent documents. The results obtained from our experiments are encouraging and are discussed in detail in this paper. We have obtained a precision of 73.2 % and a recall of 68.3%.

1 Introduction

In this work we describe a machine learning approach to extract concepts and their relations from patent documents using syntactic information. We consider patent documents for the present study because of their complex sentence constructions, both at syntactic level and semantic level. Patent documents are technical and legal documents which contain technology innovations written in legal language style. A patent document consists of several sections such as abstract, prior art, novelty or claim, the methodology and figures.

In general, a legal document comprises of long and semantically very complex sentences. When this is combined with technical writing the sentences become very complex and are difficult to be analysed. It has been observed that a single sentence in a patent document consists of more than 200 words. Hence a patent document contains sentences that are syntactically and semantically complex, which are very difficult to be analysed even by human beings.

Today there is great need for the automatic analysis of patent documents for applications such as prior art search, patent infringement. Conceptual graph (CG) (Sowa, 1984) allows representing semantic and syntactic knowledge of a sentence in a graphical model which can be used in automatic inferences. Mineau et al, (2000) observe that the conceptual modeling languages offer more expressiveness than traditional modeling languages. The extracted CGs can be transformed to corresponding first-order-logic formulae (Amati, 2000) using available tools such as COGITANT and COGUI, which facilitate further semantic inferences.

A conceptual graph (CG) is a graph representation for logic based on the semantic networks of artificial intelligence and the existential graphs of Charles Sanders Peirce (Sowa, 1984). It is a type of semantic network of concept nodes and relation nodes. Mathematically CG is a bipartite graph consisting of two types of nodes; concept nodes and relation nodes. CGs were first introduced by John Sowa in his work on database interfaces (Sowa, 1976). It illustrated CGs for representing natural language questions and mapping them to *conceptual schemata*. Each schema contained a declarative CG with attached *actor nodes* that represented functions or database relations.

Sowa (1984) explains CGs for the representation of natural language texts. The concept nodes represent entities, attributes, events, actions. And relation nodes represent the kind of relationship between the two concept nodes. The main advantage of representing natural language text in the form of CG is that, CGs can be easily converted to any Knowledge Interchange Format such as first order logic, hence semantic processing is possible. The challenge in automatic representation of a natural language text in CG is the identification of concepts and the relationships between them. Hence automatic identification of concepts and conceptual relations is very important for the purpose of semantic representation and inference.

Conceptual graphs have been used in applications such as question answering systems and information retrieval systems to improve the performance of the systems. Molla and Van (2005) build "AnswerFinder" - a framework for QA systems – in TREC 2004. Here in the graph patterns between the questions and answers is learnt. The conceptual graphs are based on translation of logical forms of sentences in the training data of question and answers given in TREC 2004. The graph matching algorithm is based on the maximal graph overlap. Here they obtain average accuracy of 21.44% and average mean reciprocal ratio of 25.97%.

Siddiqui and Tiwary (2006) use CGs for representing text for the information retrieval task. They use CG in conjunction with VSM model for representation. Here the information retrieval task is done in a two phased manner. In the first phase the relevant documents are retrieved using the VSM model. The resultant documents are used as input for the CG model and the finally most relevant documents are retrieved. Here a small set of semantic relations are used to construct CGs, these relations are developed based on the syntactic patterns. CACM-3024 data collection is used for the experiments. They show an increase of 34.8% in precision and overall 7.37% improvement in retrieval performance.

Montes-y-Gomez et al (2000, 2001) discuss about information retrieval using CGs. In this work they present methodology for comparison of two conceptual graphs. The similarity measure is based on the dice coefficient, takes into consideration both concepts and relations of the graph while calculating similarity.

Conceptual graphs are also used in developing knowledge base. Karalopoulos et al., (2004) use CGs for representing geographic knowledge. In their work, they create a CG for each geographical definition. All similarly created CGs are interconnected to form a network, thus a geographic knowledge base is developed.

The analysis we have done shows that the general in-depth parsers available in the open domain fail to parse long sentences in the patent documents. The failure of in-depth parsers led us, to develop a methodology, which uses shallow parser and NER to identify concepts and their relationships and towards development of CGs. Thus in the present work we came up with a methodology which uses shallow parsing and machine learning technique for identifying concepts and relations from sentences in the claim/novelty section of patent documents. The paper is further organized as follows: Section 2 describes the challenges or issues in the present work. Section 3, describes our approach to solve the issues encountered. Section 4, describes our experiments and results.

2 Challenges in the creation of CGs from Patent Documents

The most prominent and technically important section of a patent document is the claim /novelty section which describes and defines the claims of the invention. This describes the core novelty of the invention, for which protection is claimed by the inventors. A claim in the patent document could be classified into two types, independent and dependent claims. Independent claims introduce a unique novelty feature of the invention, whereas a dependent claim describes more about the novelty

already mentioned in an independent claim. Most of the independent claims on an average consist of 250 – 400 words in a sentence. These sentences cannot be parsed correctly by the general parsers available in the open source and also crash the system.

In (Yang & Soo, 2012), the sentences in the claim section were split into different parts using few heuristics, so that those could be parsed by a parser. The output thus obtained from a parser is used for developing CG. The problem with the sentence splitting as explained in (Yang & Soo, 2012) is that it does not retain the full meaning of the sentence and there could be information loss. Problems in such simplification for ease of processing is that of gap-filling expressions, ordering of phrases in the sentences, maintaining discourse coherence will lead to improper interpretation of sentences.

Another major challenge in the construction of CGs is defining concepts and relations between the concepts. A patent document describes main invention which can be an object or a process or a product (which again is an object). Along with this it also describes the components or parts of the object and sub-processes. It also describes the properties, characteristics, uses and advantages. In general a concept is defined as an abstract idea conceived mentally by a person. A concept will have same meaning across the languages, but have different lexical representations. A concept can be represented by a single word or a group of words. According to this definition every noun becomes a concept and identifying the relation between concepts becomes difficult. To overcome this difficulty we use concepts of the verb as well as that of the noun from WordNet (Miller, 1995) to get the relationship.

3 Our Approach

A CG consists of two types of nodes viz., concepts and relations. CGs are bipartite graphs, thus there cannot be direct linking between the same type of nodes such as concept-concept or relation-relation. Two concept nodes are connected by a relation node. Concepts are entities, attributes, states, and events such as person names, component names, phenomena, objects. Relations show how two concepts are connected with each other. For example in the sentence "*The keyboard for use with a mobile device in which a display screen for displaying output to a user is provided, keyboard has: a} a plurality of keys, each key is transparent; and b} a housing for supporting keys*". We define "Artifact: keyboard", "Artifact: mobile device", "Artifact: display screen", "Person:user", "Artifact: keys", "Attribute: transparent" and "Artifact: a housing" as concepts and relationship can be defined as "perception" (or displaying), "consumption" (or use), "stative" (or has, is), "for". A higher level semantic representation of the sentence can be "[Artifact] -> (perception) -> [Person]". We observe that relationships between two concepts (or entities) are indicated or expressed using verbs, prepositions, conjunctions in a sentence. In this work we identify the associated relationships expressed by these categories of words or phrases using CRFs.

CRFs is a supervised learning technique. CRFs have been used successfully in the past for tasks such as named entity recognition (Benajiba et al, 2009; Vijayakrishna & Sobha, 2008), semantic role labeling (Cohn et al, 2005) and cause-effect relation identification (Menaka et al, 2011). CRFs were used for sequence labeling tasks. Advantage of using CRFs is that it provides for efficient inference methodology and is tractable in the real time and it is feature based learning. Features could be any such as lexical, syntactic, and semantic. Features are easily understandable to human as well as the system.

One of the challenging issues in automatic CG construction is the identification of relations between concepts triggered by verbs. In patent documents the relations can be identified or triggered by prepositions, verbs, conjunctions. Since prepositions and conjunctions are a defined finite set of words, we can construct or derive a relation mapping ontology manually. Whereas for verbs, to derive manually a relation mapping ontology is difficult, involves lot of human effort. Apart from this we have known issues of sense ambiguity and structural ambiguity. In such a scenario use of machine learning techniques such as CRFs will be helpful and more suitable. We need to create enough and sufficient amount of manually annotated corpus which can be used for training.

The training corpus is manually tagged for the concept-relationship pair. We use the semantic information of concept classes given in the WordNet for nouns and verbs for associating relationships.

WordNet defines 25 noun concepts and 16 verb concepts. The WordNet defined concepts are for general domain and are not domain specific. The tagset used for annotating <concept-relation-concept> tuple is as follows:

a) **R-relation_type-<idx>-S** -- Here <idx> is relation index number in a sentence it is from 1,2,..N. *relation_type* specifies the type name of the relation. "S" indicates start. If relation is indicated by more than one word, then B-I-O standard format of representation is used.
b) **C1-<idx>-S** -- Here C1 refers to concept 1 of Relation R<idx>. "S" indicates start.
c) **C2-<idx>-S** -- Here C2 refers to concept 2 of Relation R<idx>. "S" indicates start.

A single concept can occur inside different relations, to accommodate this we have used the notation of indexing. As it can be seen in the above tagset, <idx> is the indexing used for this purpose. A sample annotation schema for a partial sentence is shown below.

A Portable/C2-1-S Electronic/C2a-1-S {attribs}/R-attribs-1-S appliance/C1-1-S,C1-2-S comprising/R-comprise-2-S : a keypad/C2-2-S,C1-3-S having/R-has-3-S a plurality/C1-3-S of/C1-3-I keys/C1-3-I

A tool called CRF++ (Kudo, 2005) was used for the CRFs implementation. This tool is available as open source. CRFs is an undirected graphical model, where the conditional probabilities of the output are maximized for a given input sequence (Lafferty et al, 2001). We chose CRFs, because it allows linguistic rules or conditions to be incorporated into machine learning algorithm. CRFs make a first order Markov independence assumption and can be viewed as conditionally trained probabilistic finite state automata. The training of the CRFs requires iterative scaling techniques, where a quasi-Newton method such as L-BFGs is used.

3.1 Features Used for CRFs Training

From the analysis of the patent documents we have identified the following features.
I. **Basic Syntactic features**:
a) Word, POS, chunk and Clause boundary
b) NE Tag feature
II. **Semantic features**: WordNet concept class III. **Structural features**:
a) Combination of clause boundary information, POS and chunk
b) Combination of POS and WordNet concept class
c) Combination of POS and chunk given the concept class for preceding and succeeding words
IV. **Positional features**:
a) Position index – position of the concept in the sentence
b) Clausal type – concept is in which type of clause
c) Clausal position – within a clause what is the position of the concept, this has value of phrase counts in the clause

3.2 Conceptual Graph Construction

The <concept-relation-concept> tuple obtained from CRFs is a graph. For a sentence we get several such tuples depending on the number of clauses. We merged these sub-graphs of the sentence to form a CG. Sub-graphs are merged by computing clique-sum. In this method two graphs are merged by merging them along the shared clique. A clique in a graph is a subset of vertices in which every two vertices are connected by an edge. Each tuple can be considered as a clique. We identify the shared cliques and merge them to form a single CG. For example let us consider the following patent claim sentence

(1) *"A portable electronic appliance comprising: a keypad having a plurality of keys, wherein each of the plurality of keys is arranged so as to actuate a respective mechanical switch so as to provide a first type of user input; and an impedance sensing means disposed integrally with the keypad so as to provide a second type of user input that is characterized as non-mechanical, wherein the impedance sensing means operates as a proximity sensitive touchpad, wherein the keypad and the impedance sensing means are coextensive, wherein the impedance sensing means is of a size that is always adaptable for use in a hand-held device, and wherein the impedance sensing means is disposed under the keypad."*

Here we obtain the following <concept-relation-concept> tuples as machine output (Fig. 1).

We perform the graph operation "join" to form a single conceptual graph as shown in Fig. 3.

3.3 System Architecture

The system architecture is shown in Fig. 2. The following are the modules:

a) **Claim Section Extraction:** We cull out the claims section in the patent document using heuristic rules.

b) **Preprocessing:** In this module we take the novelty, claims and advantages parts extracted using the IE module and preprocess these for part-of-speech tagging, NP-VP chunking, clause boundary tagging. And these parts are further used for conceptual graph representation.

a) [appliance] → (attribs) → -1. [portable]
-2. [electronic]
b) [appliance] →(comprise) →-1. [keypad]
-2. [impedance sensing means]
c) [keypad] → (has) → [keys]
d) [keys] ← (agent) ← [actuate]
e) [actuate] →(theme) →[mechanical switch]
f) [mechanical switch] ← (agent) ← [provide]
g) [provide] → (theme) →[user input]
h) [user input] → (attrib) → [first type]
i) [impedance sensing means] ← (agent) ← [operate]
j) [operate] → (theme) → [touchpad]
k) [touchpad] → (attrib) → [sensitive]
l) [impedance sensing means] ← (agent) ← [provide]
m) [provide] → (theme) → [user input]
n) [user input] → (attrib) → [second type]
o) [user input] → (alias) → [non mechanical input]
p) [impedance sensing means] →(attrib:size) → [adaptable] →(for) →[hand-held device]
q) [impedance sensing means] → (coextensive) →[keypad]
r) [impedance sensing means] →(under) →[keypad]

Figure 1. Concept-Relation-Concept tuples obtained as system output.

c) **Semantic Enrichment:** After preprocessing we tag the words for their concept classes from the WordNet

d) **Feature extraction:** As explained in section 3.1 the required features are culled out and represented in the format as required for CRFs training and testing.

e) **Training Phase:** During the training phase, the text is manually annotated for the <concept-relation-concept tuples>. Along with the features, training using CRFs is done. CRFs is a supervised machine learning technique.

f) **Testing Phase:** During the testing phase only the culled out syntactic, semantic and other features are presented for the CRFs engine to identify the <concept-relation-concept> tuples. Here we do not provide any manual annotation.

e) **Graph builder module:** After the <concept-relation-concept> tuple identification as explained in section 3, using the graph operation rules of join, copy, we develop a conceptual graph for each patent document. The graph operation rules are implemented using the heuristic and pattern rules. The graphs are represented internally using the linear format as shown in section 3.

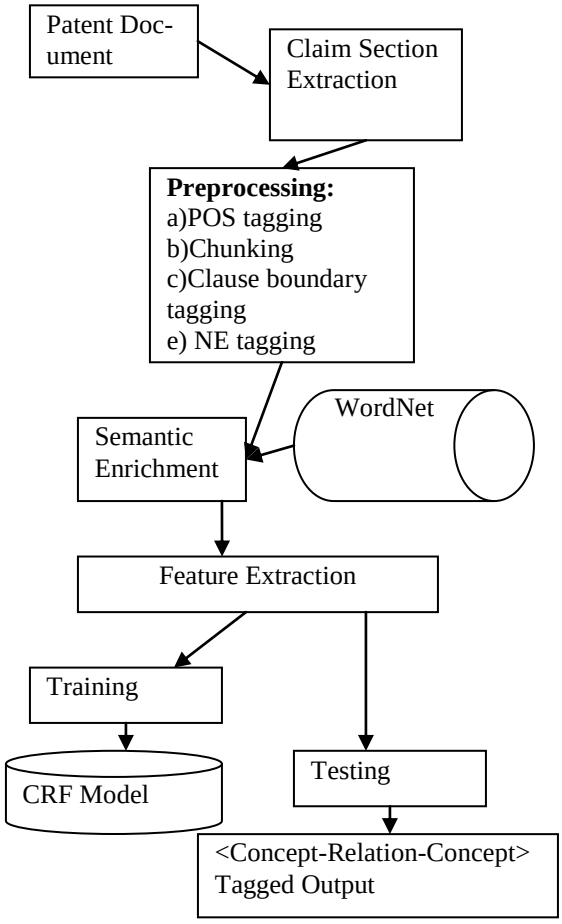

Figure 2. System Architecture

4 Experiments, Results and Discussion

In this work we have collected 1200 USPTO patent documents. The patent documents were from the domain of mobile communications, biological drug discovery and general electronics documents. The corpus consisted of 400 mobile communica-

tions domain, 300 general electronics domain, 500 from biological drug discovery domain. The patent documents were obtained from the USPTO, full-text service. The corpus was divided into two sets, training and testing.

We have performed 10 fold experiments, where we have taken 80% for training and 20% for testing. In the CRFs training we focused on identification of <concept-relation-concept> tuple for relations triggered by verbs. The evaluation metrics used are precision and recall. The average precision and recall are 73.3% and 68.3% respectively. The results of the 10 fold experiment for each individual fold is shown in Table 1.

Table1: Results of 10-fold Experiment for automatic identification of Concept-Relation-Concept Tuples

Exp. No	Total Tuples present	Tuples identified	Tuples identified correctly	Precision (%)	Recall (%)
1	7986	7458	5466	73.29	68.44
2	8050	7478	5455	72.94	67.76
3	8010	7470	5470	73.22	68.28
4	7975	7475	5469	73.16	68.57
5	8035	7460	5466	73.27	68.02
6	7965	7453	5450	73.12	68.42
7	8135	7460	5468	73.29	67.22
8	8075	7465	5463	73.18	67.65
9	7942	7467	5474	73.30	68.92
10	7926	7452	5465	73.33	68.95
AVERAGE				73.3	68.3

The analysis of the results shows that the non identification of concepts gave maximum error of 50%. In the output we observe either two concepts are combined into one or only partially identified. For example in the sentence

(2) *"an actuator assembly having an elongate plunger along a central vertical axis and having at one end a plurality of co-planar actuator surfaces arrayed initially in a horizontal plane"*

The system has combined two concepts "elongate plunger" and "central vertical axis". In some instances the concept "display screen" is identified as two different concepts "display" and "screen". This is mainly due to the pre-processing errors such as NP chunking error, where many nouns are clubbed together to form a single NP or split as different NPs. As said earlier in some instances system identifies "display screen" as two different NPs.

Though the system resolves most of the verb senses properly, in 30% of the error cases, the system does not identify the sense correctly and this has led to incorrect identification of relations. For example for the word "disposed" concept mapped from the WordNet concept is "possession", whereas the correct one would be "contact".

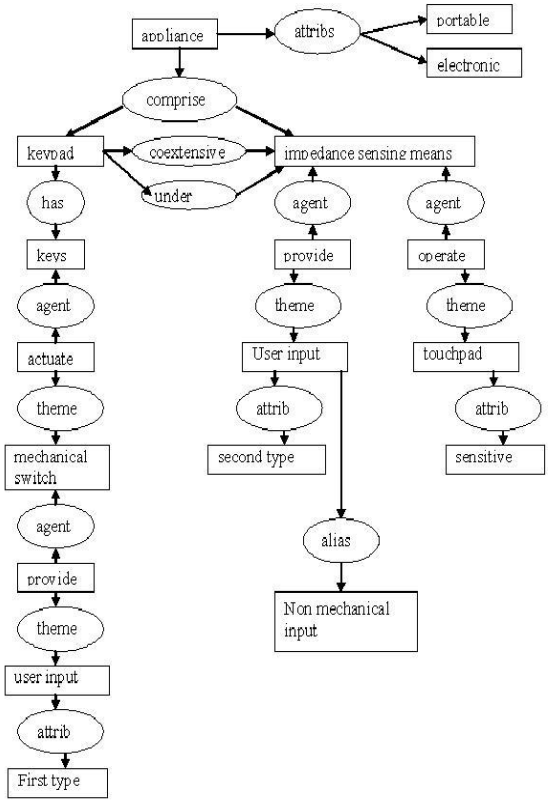

Figure 3. Conceptual Graph diagram for example (1)

5 Conclusion

The contribution of this work is in automatic identification of concepts and conceptual relations towards building CGs with the use of machine learning techniques. In this work we have used patent documents from the domain of electronics and mobile communications collected from the USPTO. The testing was done on patent documents from 2 different domains. The precision and recall are 73.3% and 68.3% respectively.

In future we plan to build a knowledge base of patent documents automatically using the conceptual graphs developed for each patent document. The CGs from each individual patent claim sen-

tence has to be merged to form such a knowledge base. We plan to use the graph comparison algorithm given by Montes (Montes et al, 2000; 2001), for this purpose. This would help in further analysing the patents.

Acknowledgments

This work is the result of the collaboration between AU-KBC Research Centre, Chennai, India and Polytechnic University of Valencia (UPV), Spain in the framework of the WIQEI IRSES project (Grant No. 269180) within the FP 7 Marie Curie.

References

Amati, G., Ounis, I. 2000. *Conceptual Graphs and First Order Logic.* The Computer Journal, 43(1):1-12.

Benajiba Y., Diab M., Rosso P. 2009. *Arabic Named Entity Recognition: A Feature-Driven Study*. In: IEEE Transactions on Audio, Speech and Language Processing, vol. 15, num. 5. Special Issue on Processing Morphologically Rich Languages, pp. 926-934.

Brill, Eric. 1994. *Some Advances in transformation Based Part of Speech Tagging*. In Proceedings of the Twelfth International Conference on Artificial Intelligence (AAAI-94), Seattle, WA.

Trevor, Cohn., Philip, Blunsom. 2005. *Semantic Role Labeling with Conditional Random Fields*. In the Proceedings of CoNLL.

Athanasios Karlopoulos., Margarita Kokla., and Marinos Kavouras. 2004. *Geographic Knowledge Representation Using Conceptual Graphs*. In the Proceedings of 7th AGILE Conference on Geographic Information Science, Greece.

Kudo, T. 2005. *CRF++, an open source toolkit for CRF*, http://crfpp.sourceforge.net.

Lafferty, J., McCallum, A., Pereira, F. 2001. *Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data*. In Proceedings of the 18th International Conference on Machine Learning (ICML-2001), pp.282-289.

Menaka S., Pattabhi RK. Rao., Sobha, Lalitha Devi. 2011. *Automatic Identification of Cause-Effect Relations in Tamil Using CRFs*. In A. Gelbukh (eds), Springer LNCS volume 6608/2011. pp 316-327.

Miller, G. A. 1995. *WordNet: A Lexical Database*. In Comm.of ACM. 38(11): 39-41.

Mineau, G., Missaoui, R., Godinx, R. 2000. *Conceptual modeling for data and knowledge management.* In Data & Knowledge Engineering, Volume. 33:137 – 168.

Deigo, Molla., Menno Van, Zaanen. 2005. *Learning of Graph Rules for Question Answering*. In the Proceedings of the Australasian Language Technology Workshop *2005*, Sydney, Australia. pp 15–23.

Montes-y-Gomez, M.., Lopez-Lopez, A., Gelbukh, A. 2000. *Information Retrieval with Conceptual Graph Matching*. In LNCS Volume. 1873, Springer-Verlag.

Montes-y-Gomez, M., Gelbukh, A., Lopez-Lopez, A., Baeza-Yates, R. 2001. *Flexible Comparison of Conceptual Graphs*. In LNCS, Volume 2113, Springer-Verlag.

Ngai, G., Florian, R. 2001. *Transformation-Based Learning in the Fast Lane*. In Proceedings of the NAACL'2001, Pittsburgh, PA. pp. 40-47.

Tanveer J. Siddiqui., Umashanker, Tiwary. 2006. *A Hybrid Model to Improve Relevance in Document Retrieval.* Journal of Digital Information Management Vol. 4(1):73-81.

John F. Sowa. 1984. *Connceptual Structures – Information Processing in Mind and Machine*. Addison Wesley.

John F. Sowa. 1976. *Conceptual Graphs for a Data Base Interface*. IBM Journal of Research and Development 20(4):336–357.

Vijayakrishna, R., Sobha, Lalitha Devi. 2008. *Domain focused Named Entity Recognizer for Tamil using Conditional Random Fields*. In Proceedings of International Joint Conference on Natural Language Processing Workshop on NER for South and South East Asian Languages, Hyderabad, India. pp. 59 - 66.

Vijay, Sundar Ram R., Sobha, Lalitha Devi. 2008. *Clause Boundary Identification Using Conditional Random Fields.* In A. Gelbukh (ed), Computational Linguistics and Intelligent Text Processing, Springer LNCS Vol. 4919/2008, pp. 140-150.

Shih-Yao Yang., Von-Wun, Soo. 2012. *Extract Conceptual Graphs from Plain Texts in Patent Claims*. Journal of Engineering Applications of Artificial Intelligence. 25(4): 874-887.