VAHA: Verbs Associate with Human Activity – A Study on Fairy Tales

Hui-Ngo Goh, Lay-Ki Soon, and Su-Cheng Haw

Faculty of Computing and Informatics, Multimedia University, Jalan Multimedia, 63100
Cyberjaya Selangor, Malaysia
{hngoh,lksoon,sucheng}@mmu.edu.my

Abstract. Named entity recognition (NER) is a subtask in information extraction which aims to locate atomic element into predefined categories. Various NER techniques and tools have been developed to fit the interest of the applications developed. However, most NER works carried out focus on non-fiction domain. Fiction based domain displays a complex context in locating its NE especially name of person that might range from living things to non-living things. This paper proposes VAHA, automated dominant characters identification in fiction domain, particularly in fairy tales. TreeTagger, Stanford Dependencies and WordNet are the three freely available tools being used to identify verbs that are associated with human activity. The experimental results show that it is viable to use verb in identifying named entity, particularly in people category and it can be applied in a small text size environment.

Keywords. Named entity recognition, fairy tales, verb, dominant character.

1 Introduction

The concept of named entity recognition (NER) is not new in the area of information extraction (IE). It has been 21 years since the first NER published which focused on extracting company names [1] using heuristics and hand-coded rules. Thereafter, various predefined categories of NER have been explored to fit the interest of the applications intended to be developed. Among all, "name of people, organzaition and location"[2], [3], [4] are the most commonly explored predefined categories. Generally, NER aims to locate and extract significant atomic elements in texts into predefined categories.

Casey *et al.* employed machine learning approach to extract multiple NEs ranging from high level (place, person) predefined categories to low level (soccer player, universities) predefined categories in the web environment. A set of seed entities and relations, and learn templates are used to automatically generates training data [5]. Einat *et al.* applied conditional random fields and dictionary to extract personal names from email [6]. However, the unstructured nature of written email produced inconsistent performance results among corpora used. Repetition of names [7], [8] within documents have also been used to extract NE, but it might not perform well for documents that are small in size.

H. Jiang et al. (Eds.): IEA/AIE 2012, LNAI 7345, pp. 313–322, 2012.

In 2010, Le *et al.* studied the use of inductive logic programming to extract named entities (name, diploma, organization, research) in Vietnamese language [9]. In the same year, Laura *et al.* proposed domain adaptation of rule-based annotator to enhance domain customization for NER by manually coded 104 features of domain-independent CoreNER library [2]. Public datasets of CoNLL03, Enron and ACE05 were used to train and test the "person, location and organization" entities.

However, most NERs developed above focused on non-fiction based documents. Non-fiction implies communicative works whose descriptions are generally written as facts. Therefore, non-fiction documents usually exhibit certain patterns in representing its NE. For an instance, name of person may start with designator, capital letter of the first character, and naming in a human way. On the other hand, fiction documents usually exhibit complexity and uncertainty in locating its NE. For example, the name of a person may be represented in diverse spectrums, ranging from living things (animals, plants, person) to non-living things (vehicle, furniture).

In this paper, we propose VAHA, a fully automated named entity recognition framework to overcome the above mentioned issue by studying the nature of verb(s) that associates with human activity. We aim to extract dominant characters in fairy tales. Dominant characters are the person depicted in a narrative and actively engage to get audience attention. Usually they are clearly identified through an impact play in a story regardless of the life span of their appearance in the story. Fiction-based domain is used to test the proposed framework. A predefined category of "name of person" is being investigated but our approach focuses on recognizing dominant characters in fairy tales. Stanford dependencies (SD) and TreeTagger are used to shallowly parse the natural language input files to identify the potential dominant character(s). Clauses that are tagged with the sequences of (i) noun(s) - verb(s) denotes Subject-Verb (S-V) and (ii) verb(s) - noun(s) implies Verb-Object (V-O) are being extracted. The extracted S-V and/or V-O at sentence level will then be verified with semantic dependencies produced by SD to conform it represents the sentence meaning. Later, two features of WordNet, namely derivationally related form (DRF) and sentence frames (SF) are used to substantiate verb that associates with human activity. Finally, subject or object that attach to the verb that associates with human activity will be regarded as dominant character. Part of this work is an extension of our previous effort in identifying protagonist in fairy tales using verb. This paper is segmented into 3 sections of technologies background, proposed system framework, experiments performed and end with conclusion.

2 Technologies Background

2.1 TreeTagger

TreeTagger[1] is a tool developed within the TC project at the Institute for Computational Linguistics of the University of Stuttgart for annotating text with

[1] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

part-of-speech and its lemma information. It is readily available in eight languages and adaptable to other languages with the availability of the training corpus. In this work, English is used to annotate on the selected natural language text file.

```
Between/IN these/DT pieces/NNS grew/VVD the/DT most/RBS beau-
tiful/JJ large/JJ white/JJ flowers/NNS;/: so/RB the/DT swal-
low/NN flew/VVD down/RP with/IN Tiny/NP,/, and/CC placed/VVD
her/PP on/IN one/CD of/IN the/DT broad/JJ leaves/NNS./SENT
```

Fig. 1. Natural language sentence with its corresponding TreeTagger tag

Fig. 1 shows the sentence that has been annotated with TreeTagger. Each word is attached with its corresponding part-of speech (POS) tag. An assumption is formed where dominant character(s) of a fairy tale is often tagged as "NOUN" generally, and specifically as common nouns and/or proper nouns. In this work, "NOUN" is identified as any POS tags that start with the label "N" such as "NN", "NNS" or "NP" whereas "VERB" is POS tags that start with the label "V". Clauses that are tagged with the sequences of (i) noun(s) - verb(s) denotes Subject-Verb (S-V) and (ii) verb(s) - noun(s) implies Verb-Object (V-O) are extracted. As shown in Fig. 1, there are two S-V extracted, namely "pieces/NNS grew/VVD", "swallow/NN flew/VVD".

2.2 Stanford Parser

2.2.1 Stanford Dependencies

Stanford Dependencies (SD) formulates its dependencies relation based on forty-eight grammatical relations according to the predefined tregex patterns over phase-structure trees [10]. Tregex is a matching patterns in trees based on tree relationship and regular expression. SD is represented in triplet structure with a grammatical relation used to tie up the right dependencies of two tokens as shown in Fig. 2.

In VAHA, SD is used to filter out each extracted S-V and V-O that does not conform to its corresponding sentence meaning. Grammatical relation of "nn", "nsubj", "nsubjpass" or "dobj" are used to examine against each extracted S-V and V-O at sentence level. Only S-V and V-O that have its corresponding SD with any of the four mentioned grammatical relations will be kept for further analysis. "nn" denotes noun compound modifier that serves to modify the head noun, "nsubj" implies nominal subject which is the syntactic subject of a clause whereas "nsubjpass" refers to passive nominal subject which is the syntactic subject of a passive clause; "dobj" means direct object of a Verb Phrase (VP) which is the object of a verb.

```
det(pieces-3, these-2)
prep_between(grew-4, pieces-3)
det(flowers;-10, the-5)
advmod(beautiful-7, most-6)
amod(flowers;-10, beautiful-7)
amod(flowers;-10, large-8)
amod(flowers;-10, white-9)
nsubj(grew-4, flowers;-10)
mark(flew-14, so-11)
det(swallow-13, the-12)
nsubj(flew-14, swallow-13)
dep(flowers;-10, flew-14)
prt(flew-14, down-15)
prep_with(flew-14, Tiny,-17)
conj_and(flew-14, placed-19)
dobj(placed-19, her-20)
prep_on(placed-19, one-22)
det(leaves.-26, the-24)
amod(leaves.-26, broad-25)
prep_of(one-22, leaves.-26)
```

Fig. 2. Stanford dependencies

Fig. 2 presents the grammatical dependencies for the sentence in Fig. 1. Based on the two extracted S-V mentioned in section 2.1, “pieces/NNS grew/VVD” does not conform with the sentence meaning as its grammatical relation is none of the four mentioned relation. It is not the “pieces” that grew, but in fact, it is the flower that grew. Hence, “pieces/NNS grew/VVD” will be discarded for its verb analysis. However, S-V of “swallow/NN flew/VVD” will be kept for further verb analysis as it conform to the sentence meaning and has the grammatical relation of “nsubj”.

2.3 WordNet

WordNet [11] is an English lexical database that group set of words into Synonyms (Synsets). Each synset is interconnected by conceptual relations. Hence, it illustrates the co-reference among synsets in database in revealing the semantic represented. As of 2006, the WordNet database contains total of 155,287 unique words where verb has taken up 11,529 words organized in 13,767 synsets for a total of 25,047 word-sense pairs [12] which is sufficient to be used in this project. Two features of WordNet are used in this work, namely derivationally related forms (DRF) and sentence frames (SF). DRF indicates words that are derived from the same root. It shows relationship existed between groups of synsets. SF is specifically designed for VERB group; it contains a list of generic sentence frames exemplifying the types of simple sentences in which the verbs in the synset can be used.

3 Framework and Experimental Setup

This section explains the framework as well as the experimental setup.

Input : Fiction web pages[2] which contain eight fairy tales, as listed in Table 1 are chosen as they contain diverse spectrum of dominant character(s). Some of the dominant characters are represented as its real name like human being while some of the dominant characters are being symbolized as animal or inserts.

[2] http://www.kidsgen.com

Table 1. Fairy tales with the corresponding word count

Fairy Tale	Word Count
The Story of Snow White	1913
Cinderella	1077
Beauty and the Beast	1357
Rapunzel	1393
Thumbelina	4348
Ugly Duckling	841
Sleeping Beauty	1317
Ant and the Grasshopper	142

Step 1: Document cleaning
Each fairy tale web page is cleaned automatically using HTML Context Extractor [3]in order to get rid of non-text content (banner, audio, video, images). A pure text file (.txt) is produced at the end of the cleaning process.

Step 2: Pre-linguistic processing
Each pure text file will be shallowly parsed using two freely available text processing tools, namely (a) TreeTagger and (b) Stanford Parser.

(a) The POS tags annotated on the pure text file served as a scheme in extracting clauses (S-V and/or V-O) that contain potential dominant character(s).
(b) Grammatical relations supplied by Stanford dependencies (SD) is used to verify that the extracted patterns (S-V and V-O) conform with its corresponding sentence meaning.

Step 3: Feature extraction
Based on each annotated fairy tale using Treetagger, potential dominant characters are extracted based on two patterns, namely (a) Subject-Verb (S-V) and (b) Verb-Object (V-O).

(a) For S-V pattern, clause that contains noun(s)-verb(s) and noun(s)-who-verb(s) that are adjacent to each other will be extracted.
(b) For V-O pattern, determiner might appear in between verb and object. It is an article which is to introduce a noun. "a", "an" and "the" are the examples of article. Therefore, clause that contains verb(s)-noun(s) or verb(s)-determiner-noun(s) that are adjacent to each other will be extracted.

Step 4: Data filtering
Two filtering processes of (a) conformation of extracted S-V and V-O with its corresponding sentence and (b) main verb identification are performed to prepare an accurate data for verb analysis.

(a) Four grammatical relations of "nn", "nsubj", "nsubjpass" and "dobj" from SD are used to countercheck against all the extracted S-V and V-O. It aims to filter out unconformation of S-V and V-O with its corresponding sentence meaning. An example is illustrated in section 3.2.

[3] `http://senews.sourceforge.net/KCE_README.html`

(b) Given the filtered S-V and V-O from step 4(a), only verb which has POS tag of "VV", "VVD", "VVG, "VVN", "VVP" and "VVZ" are preserved for verb analysis. These POS tags have the main verb that describes an event taken by a subject or action imposed on an object, such as "take/VV", "took/VVD", "taking/VVG", "taken/VVN", "take/VVP" and"takes/VVZ". Section 2.1 describes TreeTagger POS tags.

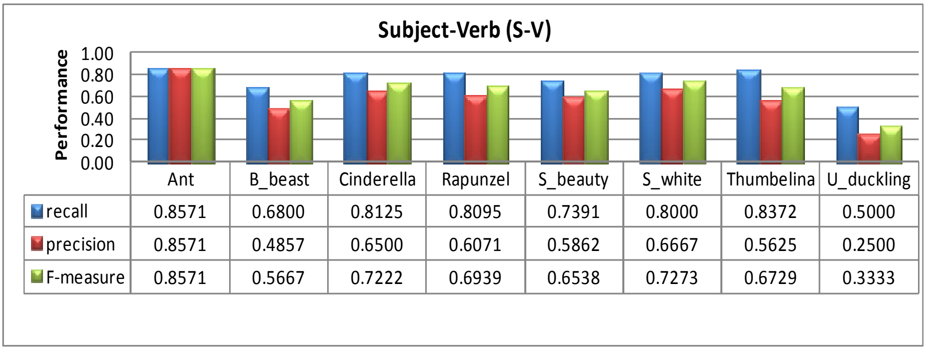
Finally, a set of filtered clauses is produced. They are notated as S-V_{can} and V_{can}-O.

Step 5: Verb analysis
Each extracted V_{can} that forms S-V_{can} or V_{can}-O will be served as keyword search in WordNet for retrieving its corresponding senses description. Only V_{can} that has the returned group(s) of VERB and /or ADJECTIVE from WordNet will be regarded for its DRF or SF. For DRF, each returned description will be examined sentence by sentence. In the presence of either key phrases of "someone who", "a person who", "to whom" and "one who" in the sentence, the verb is considered to be associated with human activity. However, in the absent of four key phrases in DRF, sentence frames is used to study the verb's usage. For S-V, if all the V_{can}'s sentence frames start with "Somebody ---- ", it implies an action has to be taken by human. For an instance, the word "sighed" for S-V of "Cinderella/NP sighed/VVD" in the story of "Cinderella". While for V-O, if all the V_{can} sentence frames end with "---- Somebody", it denotes an action is taken on human. Hence, the word is considered to be associated with human activity.

Output: Finally, S and O which are attached to verb(V) that associates with human activity are considered as dominant characters in the investigating fairy tale.

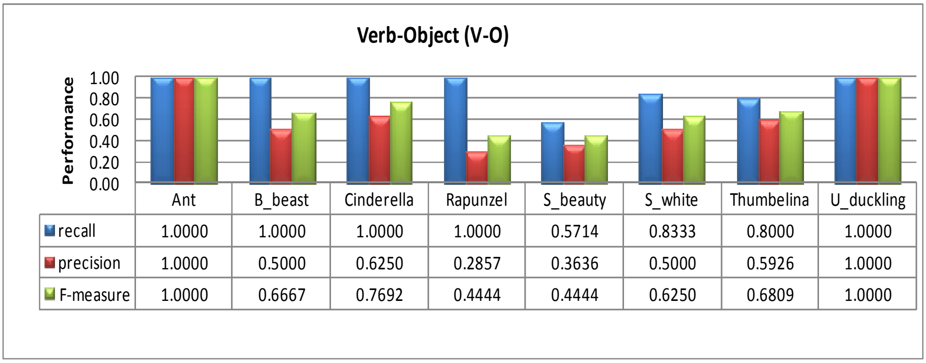
4 Results and Discussion

The eight chosen fairy tales that come in different file sizes and have diverse dominant characters are experimented to verify the performance of our proposed VAHA in identifying dominant character(s). The evaluation metrics used are precision, recall and F-measure. The experimental analysis were carried out individually for S-V as depicted in Fig. 3 and V-O as presented in Fig. 4, while the overall performance of VAHA is shown in Fig. 5.



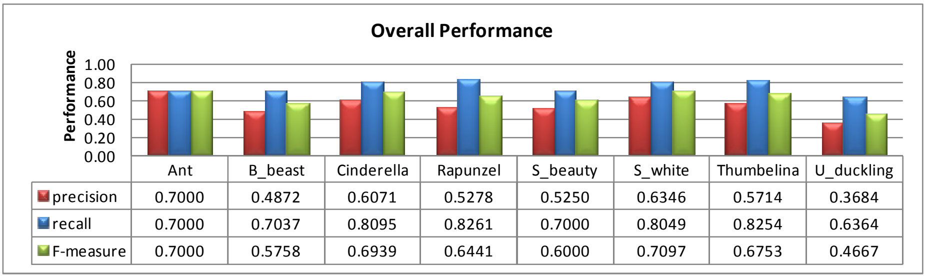
	Ant	B_beast	Cinderella	Rapunzel	S_beauty	S_white	Thumbelina	U_duckling
recall	0.8571	0.6800	0.8125	0.8095	0.7391	0.8000	0.8372	0.5000
precision	0.8571	0.4857	0.6500	0.6071	0.5862	0.6667	0.5625	0.2500
F-measure	0.8571	0.5667	0.7222	0.6939	0.6538	0.7273	0.6729	0.3333

Fig. 3. Performance results of "S-V" pattern for dominant character(s) identification



	Ant	B_beast	Cinderella	Rapunzel	S_beauty	S_white	Thumbelina	U_duckling
🟦 recall	1.0000	1.0000	1.0000	1.0000	0.5714	0.8333	0.8000	1.0000
🟥 precision	1.0000	0.5000	0.6250	0.2857	0.3636	0.5000	0.5926	1.0000
🟩 F-measure	1.0000	0.6667	0.7692	0.4444	0.4444	0.6250	0.6809	1.0000

Fig. 4. Performance results of "V-O" pattern for dominant character(s) identification



	Ant	B_beast	Cinderella	Rapunzel	S_beauty	S_white	Thumbelina	U_duckling
🟥 precision	0.7000	0.4872	0.6071	0.5278	0.5250	0.6346	0.5714	0.3684
🟦 recall	0.7000	0.7037	0.8095	0.8261	0.7000	0.8049	0.8254	0.6364
🟩 F-measure	0.7000	0.5758	0.6939	0.6441	0.6000	0.7097	0.6753	0.4667

Fig. 5. Performance results for overall dominant character(s) identification

As shown in Fig. 5, file size and dominant characters' groups do not impact the performance results of dominant character identification. This can be validated through the story of "Ant and Grasshopper" which has file size of only 142 words with the character group of inserts yields the best performance result compared to other fairy tales. Out of seven clauses of "*grasshopper* – was hopping", "*ant* - passed", "*grasshopper* – invited", "*ant* - sit", "*ant* -went" "winter-came" and "*ants* - distributing" being extracted for S-V pattern, six of the subjects (highlighted in italic) were correctly classified as character based on the verb attached to them except for the subject of "winter". While, "asked – the ant", "invited – the ant" and "said – the grasshopper" are the three clauses extracted for V-O pattern. The word of "asked" and "invited" are related to human activity. However, an observation was done for the word "said" of V-O pattern in all the investigated fairy tales. V-O clause that has the pattern of "verb-determiner noun" and appear immediately after the punctuation mark of " " " or " , " always denote noun as character except for one clause of "said – the spirit" in the story of " Snow White". "replied – the field mouse" and "exclaimed - the field mouse" are the two examples that share the same characteristics of the word "said" in the story of "Thumbelina". Therefore, clause that exhibits the above characteristics for V-O pattern are taken as a heuristic in identifying dominant characters.

The analysis for S-V pattern is rather straightforward as DRF exhibits its description of key phrases of "someone who", "a person who", "to whom" and "one who" in the form of active sentence. For example, “someone who consumes”, implies subject must be a person, while, V-O pattern possessed a more complex situation where an action (verb) can be taken on a person or thing. Besides, “drank – the dew”, “heard – a voice”, “spun – gold” and “sing – a wedding song” are some of the V-O extracted from fairy tales. As such, a correspondence senses between DRF and SF is needed to verify the uncertainty. SF that ends only with “---- something” implies the said object must be a thing and SF that ends only with “---- somebody” implies the said object must be a person. However, for the cases where SF that has the mixture of “something/somebody” at subject and/or object are currently ignored for this work and will be explored in future work. Therefore, the performance result of V-O pattern (Fig. 4) generally performs better than S-V pattern (Fig. 3). This is due to the number of clauses extracted for S-V pattern is more than the V-O pattern. Hence, the possibility of subject wrongly classified as character is higher. Moreover, structural format of V-O pattern has led to more details handling compared to a straightforward S-V pattern.

Recall, precision and F-measure are interrelated. Good information extraction should reflect high recall and high precision for high F-measure, which is hard to be achieved. High recall with low precision or low recall with high precision always becomes a struggling effort for researcher. An effort to improve either factor might cause the other factor to be deteriorated. In this work, high recall implies most of the verbs that associate with human activity are in fact attached to dominant character while low precision is due to many verbs which are not associated with human activity were being extracted as according to the patterns of S-V and V-O. Hence, the number of extracted pattern has increased and greatly impacts the result for precision. This scenario can be seen in Fig. 3, 4 and 5.

Table 2. Dominant dominant characters for fairy tales

Fairy Tale	Dominant Character
The Story of Snow White	**Snow White, King, Queen, Stepmother**(witch**, peddler woman), Prince, Huntsman,**
Cinderella	**Cinderella, Prince, Coachman**, **Stepmother**, Stepsister, **Fairy, Minister**.
Beauty and the Beast	**Merchant, Beast, Beauty.**
Rapunzel	**Enchantress, Rapunzel, Prince, Husband,** Wife.
Thumbelina	**Tiny, Toad, Swallow, Mole, Field Mouse, Cockchafer, Butterfly, Bird**, Prince.
Ugly Duckling	**Mother duck, Duckling, Geese, Hen**, Swan, Farmer
Sleeping Beauty	**Briar Rose**, Witch, **King, Queen, Frog, Prince**.
Ant and the Grasshopper	**Ant, Grasshopper**

Five different versions of each fairy tale were used to identify the common dominant characters. The existence of dominant characters in at least three versions the same fairy tale will be chosen as true dominant character. Table 2 presents the true dominant characters for each fairy tale. Dominant characters which are highlighted in bold are the dominant characters identifiable using VAHA. The repetitive occurrence of dominant character in subject and object is counted as one occurrence. In the stories of "The Story of Snow White", "Beauty and the Beast" and "Ant and Grasshopper", our approach is capable to identify all of the listed dominant characters. However, VAHA did show good performance too in the rest of the stories. The active participation of each dominant character in story flow will likely increase its affiliation to verb that associates with human activity. An example of S-V and V-O patterns for "Ant and Grasshopper" shown above explain this justification. The dominant character "stepsister" is unidentifiable in the story of "Cinderella" as there are only two S-V pattern of "stepsister – getting", "stepsister – gaped". Moreover, the verbs of "getting" and "gaped" are not associated to human activity.

5 Conclusion

This paper describes VAHA, an algorithmic framework for automatic identification of dominant characters in fairy tales by studying the nature of verb that associates with human activity. Two different groups of dominant characters were used to test on VAHA, namely, entity and human alike name of dominant characters. TreeTagger, Stanford Dependencies and WordNet are the three freely available tools being used to identify verb that associates with human activity. Different handling has been taken on S-V and V-O pattern due to the different structural representation and characteristics of DRF exhibits. Our experimental results show that verbs can be used as a determinant in identifying "people" named entity in general and protagonist in specific. For future work, we wish apply VAHA in news articles and to look into verb disambiguation to improve the performance result of dominant characters identification.

References

1. Lisa, F.R.: Extracting Company Names from Text. In: Proc. IEEE Conference on Artificial Intelligence Applications, pp. 20–32 (1991)
2. Laura, C., Rajasekar, K., Li, Y.Y., Frederick, R., Shivakumar, V.: Domain Adaptation of Rule-Based Annotators for Named-Entity Recognition Tasks. In: Empirical Methods in Natural Language Processing, Massachusetts, pp. 1002–1012 (2010)
3. Andrew, M., Li, W.: Early Results for Named Entity Recognition with Conditional Random Fields, Feature Induction and Web-Enhanced Lexicons. In: 7th Conference on Natural Language Learning, pp. 188–191 (2003)
4. Dan, K., Joseph, S., Huy, N., Christopher, D.M.: Named Entity Recognition with Character-Level Models. In: 7th Conference on Natural Language Learning, pp. 180–183 (2003)
5. Casey, W., Alex, K., Nemanja, P., Lyle, U.: Web-Scale Named Entity Recognition. In: Proceedings of the 17th ACM Conference on Information and Knowledge Management, Napa Valley, California, USA, pp. 123–132 (2008)
6. Einat, M., Richard, C.W., William, W.C.: Extracting Personal Names from Email: Applying Named Entity Recognition to Informal Text. In: Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, Vancouver, pp. 443–450 (2005)

7. Charles, S., Andrew, M.: Collective Segmentation and Labeling of Distant Entities in Information Extraction. In: ICML Workshop on Statistical Relational Learning (2004)
8. Razvan, C.B., Raymond, J.M.: Relational Markov Networks for Collective Information Extraction. In: ICML- Workshop on Statistical Relational Learning (2004)
9. Le, H.T., Nguyen, T.H.: Name Entity Recognition using Inductive Logic Programming. In: Symposium on Information and Communication Technology, Vietnam, pp. 71–77 (2010)
10. Marie-Catherine, D.M., Bill, M., Christopher, D.M.: Generating Typed Dependency Parses from Phrase Structure Parses. In: LREC (2006)
11. Christiane, F.: WordNet:An Electronic Lexical Database. MIT Press, Cambridge
12. WordNet Statistic, http://wordnet.princeton.edu/wordnet/man/wnstats.7WN.html