Int. J. Advance Soft Compu. Appl, Vol. 6, No. 3, November 2014
ISSN 2074-8523

Development of Machine Learning based Natural Language Processing System

Jaeho Lee[1], Younhee Kim[2], Hyunkyung Shin[3], and Kibong Song[4]

[1]Department of Computer Education, Gyeongin National University of Education
e-mail: jhlee@ginue.ac.kr
[2]Department of e-Business, Bucheon University
e-mail: yhkim@bc.ac.kr
[3]Department of Mathematical Science, Gachon University
e-mail: hyunkyung@gachon.ac.kr
[4]Electronics and Telecommunications Research Institute
e-mail: kbsong@etri.re.kr

Abstract

For disease diagnostic knowledge base system (including Q&A, consistency checker, and informativity checker) requiring higher degree of strictness on the results of a query to the system, at the development stage, great efforts are necessary to improve machine learning based statistical performance tests, measured by precision and recall rates, on the training error and prediction. Performance of the test runs is generally dependent on the two basic factors as the following: first of all, the underlying technique of in-depth context analysis on corpora with inference capability, secondly, that of user's query sentence analysis with inference capability. More importantly, a disease diagnostic knowledge base system should be able to update effectively the latest research achievements in timely manner. To meet the requirements, we propose an automatic system for construction of knowledge base from the academic archive of medical literatures. For the purpose of presentation in this paper, a prototype of knowledge base construction using natural language processing system for early diagnosis of Alzheimer disease has been designed and implemented. Since there are plenty of knowledge base systems available for Alzheimer diagnosis in English language, to differentiate our works with the existing data, we performed our research with the literatures written in Korean. The natural language processing system proposed in this paper consists of 8 modules most of which are machine learning trainer/prediction model based on maximum entropy algorithm. Tests showed that, for

all the modules, iterative training have been succeeded with precision over 90%.

Keywords: *disease diagnosis system, knowledge base system, machine learning, natural language processing*

1 Introduction

Information on science, engineering, politics, economic markets, and social networks are updated continually and archived in form of text format. Social sensor of information is necessary as Gleick says "the flood of information exposed is new challenges, as we retain more of information, it takes much more effort to delete or remove unwanted information than to accumulate it" [15]. Such information flood influences on conventional knowledge engineering [16] in term of revision cycle period for its knowledge-based system. An automation process of converting from text archives to database has long been sought in computer science [17]. The most challenging parts of that process are in the stage of conversion from natural language to logic formulation adopted in knowledge base [18]. Natural language process is also crucial to understand truthful intention from user's query sentence [13]. Especially for the case of disease diagnostic system which requires very high level of strictness on update corpora in terms of truthfulness of the information for knowledge base and accurate analysis on query sentence is more critical issue. To meet these requirements, we propose an automatic system for construction of knowledge base by using natural language processing. The system requires several standard facilities listed as follows: storage for knowledge data; language for knowledge representation and inference including predicate calculus; logical formalism such as PL (Prepositional Logic) for SAT (SATisfiability), FOL (First Order Logic) or DL (Descriptive Logic); natural language interpreter, and so on. Among those facilities, in this paper, our attention is focused on details of the natural language processing system part.

In this paper we present a proposed model design of natural language processing with the archived text data and the research results applied to the specified domain of Alzheimer disease related articles. As a language Korean has different grammar structure. For an example, the most significant discriminations are as follows: in English VP is typically consisted of the nested VP + NP, on the contrary, in Korean VP is typically consisted of the nested NP + VP; extremely hard analyze for the underspecified quantifiers due to rare use of delimiter; rare uses of the pronouns, where NP and VP denote for noun phrase and verb phrase, respectively. For the presentation purpose, our results are focused on the study with the domestic articles written in Korean language in order to avoid advantage of using the existing well organized medical knowledge base system for English language [19].

This paper is organized as follows: in section 2 the related research works are summarized; in section 3 we present the details of our design and implementation on machine learning technique based natural language processing system; in section 4 parts of experimental results are presented; in section 5 some remarks on conclusion are shown.

2 Related Works

Corpus based computational natural language processing (NLP) is a statistical approach incorporated with machine learning techniques [5, 6, 7, 8, 9, 10, 11] and one of the most active field area. Milliards of corpora are present in various academic areas (from literature to medical science) and in various languages, which are well summarized as shown at a research laboratory in Stanford university [20]. Among the many applications of NLP including machine translation, computational semantics, knowledge representations, machine learning techniques have been proactively developed with information extraction [21]. Computational semantics can be said as the most direct and prominent research area in NLP [13], in which semantic disambiguity of quantifier sentences. Among other theories, dependency tree semantics is interesting [22].

In medical science community, innovative studies on the new diagnoses for various diseases are always active and numerous experimental results are produced in form of research papers. This is flood of information cumulated into database server. On the other hand, the users of medical information require high degree of quality on their query results reflecting the latest research works [14]. In an effort to meet this requirement, techniques on information extraction using NLP from the literature have been developed progressively [1, 2, 3]. Automatic construction of knowledge base directly from the archive of literatures has been attempted since 1970’s without much success [13].

Grammar of Korean language is structurally different with that of English, which results in incompatibility of NLP when it was trained by different language and by different Tree bank. There are several of the tree bank rules currently available: namely, “Sejong Treebank” [23], as a part of “UPENN Treebank” [12], and so on. The availability of corpora in Korean is lower at this point, certainly no availability of medical corpus on Alzheimer Disease in Korean. This requires that we should build new corpora written in Korean language as well as associated probing routines for the training procedures.

Ontology technique based approaches typically make use of procedural knowledge or object-pair with interactions which are too limited formalization to be used as knowledge base for the generalized knowledge processing such as diagnosis of disease. Refolo at el. [25] recently presented an ontology structure with a few hundreds of the topics based on the seven top level categories in Alzheimer disease. Krotzsch at el. [26] contributed a new aspect of web ontology language into the ontology technique with the improved description logic.

3 Machine Learning Technique Based Natural Language Processing System

In the design of natural language system, as described at the previous section, the corpus based machine learning technique has shown its great strength when scale of text data is large. In this paper we also adopt machine learning technique for NLP, where our researches are involved with the two main works: first of all, design of the model for detectors and classifiers; secondly, design of the trainers; thirdly, construction of the training data set and the ground-truth data set. In this section, we describe that, for the first part, the eight of the classifier models: sentence detector; tokenizer; name finder; POS tagger; stemmizer; chunker; parser; co-reference resolution, for the second part, the trainer models with the maximum entropy (MaxEnt) base and hidden markov model (HMM), for the third part, details on construction of training data set.

For the specific case of study on knowledge base system on the early diagnosis of disease, both accuracy and reliability of the extracted knowledge information are highly critical. In the case of our study based on the machine learning techniques, we set the highest priority on acquisition of a ground truth corpora from the strictly peer reviewed literatures. We collected the 522 numbers of published papers on Alzheimer disease placed in the internet archives. In an effort on construction of automated knowledge base system, to minimize human interaction with the data, we set an OCR engine which converts data format from PDF to text. These text-converted data was used to build corpora for machine learning based natural language processing.

For the general NLP classifier models, such as sentence detector, Tokenizer, NER, POS Tagger, Chunker, Parser, we adopted ASF (Apache Software Foundation) OpenNLP library[4] as a basis of our NLP engine.

Fig. 1 shows an overview of our design concept for a new knowledge base system consisted of ‘continual data acquisition’, ‘NLP’, and ‘knowledge representation’. The engine of knowledge representation was constructed using the frame system based on the knowledge machine [24]. In this paper, for the presentational purpose, our discussion is weighed on the NLP module.

As can be seen Fig. 1, in this study, NLP is consisted of the eight independent modules. Each of the modules has been equipped with the trained corpora. The list of modules is as follows: sentence detector; tokenizer; name finder; POS (Part Of Speech) tagger; stemmizer; Treebank chunker; Treebank parser; co-reference resolution.

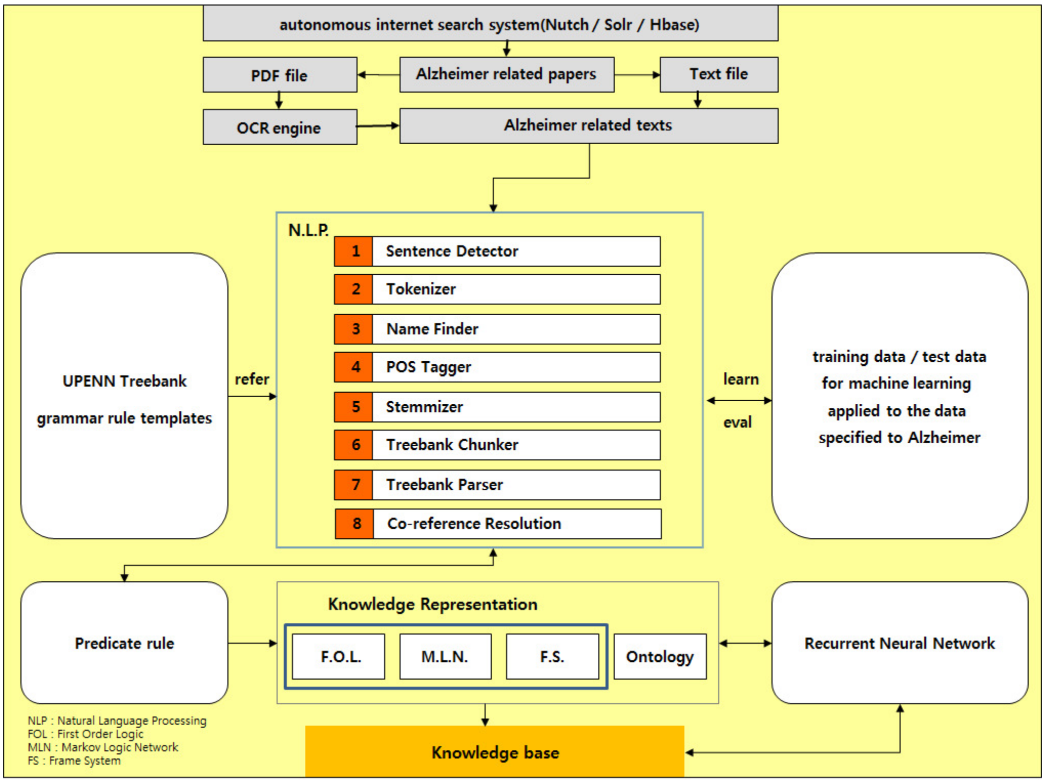

Fig. 1 Overview of knowledge base system

3.1 Sentence Detector

Like the conventional methods our sentence detector finds the patterns of boundary of sentences such as the marks of period and space. We additionally use the classifier model with supervised learning based on MAXENT (maximum entropy) [5, 6, 7, 8, 9, 10, 11]. The training data for supervised learning was built from 522 pieces of papers related with Alzheimer disease. Machine learning based classifier for sentence detection has been suffered due to limitation in performance for the specific usages cases such as period appeared in date-time and abbreviation. We resolved this issues by using rule based classifier.

3.2 Tokenizer

We, in this paper, defined the token as one of the following objects: word; sentence marks; numbers. Additionally, we refined the token in Korean language as follows: for a Korean word consisting of a stem and a group of suffixes, we separate the stem and the suffixes into different objects of tokens. Identification of tokens from sentences has been performed by MAXENT based machine learning classifier. Construction of training data for tokenization model, we put the white space to split the words into the tokens. In case there is no available space (e.g. a position between a stem of word and a suffix), we put a special tag mark '<SPLIT>'.

3.3 Name Finder

In terms of performance rates, name finder is the most critical module of natural language processing when the system is specialized for a specified subject (e.g., Alzheimer disease in our case). The tokens, obtained from the tokenizer module, are the inputs for named entity recognition engine in which a dictionary for specialized words is defined.

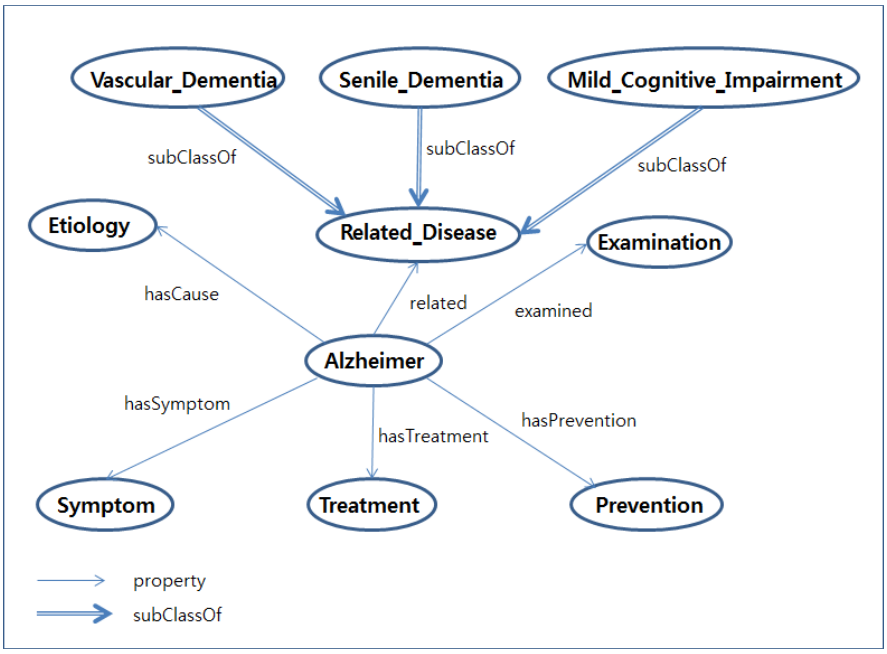

Fig. 2 A partial result of ontology modeling for Alzheimer disease

As seen in Fig. 2, in this paper, for the construction of named entity dictionary, we used the ontology structure on the terminologies and the conceptualization of relations between the terminologies. This approach method enables standardization of terminology and enhances the performance rates in inference and sentence analysis. In Fig. 2 we presents partially an ontology model for conceptualization of named entities with their relationships within the context of knowledge space related to Alzheimer disease.

For the study of this paper, we used the 7 of conventional named entity dictionary such as 'date', 'place', 'money', 'organization', 'mathematical symbol', 'person's name', and 'time'. In addition to the dictionary, we added another dictionary specialized to the medical terminology.

3.4 POS Tagger

Using the tokens identified by the tokenizer and the named entities obtained from the name finder described above, respectively, the POS Tagger module classifies them into the eight different top level classes and the 26 low level sub-classes. Unlike the sentence detector, tokenizer, name finder, POS tagging is sensitively dependent on the structure of Korean language. In this study, we adopted HMM (Hidden Markov Model) based machine learning techniques [5, 6, 7, 8, 9, 10]. For the tag sets for POS, we adopted UPENN Treebank for Korean language [12].

3.5 Stemmization

To achieve parser functionality of natural language processing on the sentence in Korean words, it is necessary to classify structures of the suffixes in Korean words. In this study, for the classification of suffixes, we used the results of POS tagging. Upon completion of building a suffix dictionary, we use regular expression based dictionary matching function for stemmization of Korean words.

3.6 Chunker

As the same way as practiced in the conventional NLP, we used the chunker model to group the words consisting in a sentence into the semantic units in form of phrase. In this paper, we concentrated our effort on providing the methods to categorize structures of the verb phrase and the noun phrase for Korean language, which is much different with English in grammar. A standard structure of phrase in Korean language is as follows: NP = N (Subj) + [N (Obj) + Adv + Adj + V]; VP = [N (Subj) + N (Obj) + Adv + Adj] + V, where the bracket notation '[ ]' indicates optional.

3.7 Parser

The parser performs syntax analysis of sentence in Korean language following the pre-defined UPENN Treebank rules. The chunking parser and CFG (Context Free Grammar) based tree-insert parser was adopted in this paper. Rather than introduction of a new Korean parser model, our studies were more weighted on the semantic analysis is based on the syntax analysis through discourse representation theory.

3.8 Co-reference Resolution

Contrasting to English language in which frequent use of the pronouns requires co-reference resolution, Old Korean language does not use the pronouns. However, use of pronoun is grammatically correct and its usage is increasing in Korean. We developed a simple module for co-reference resolution for Korean language, which are mainly used in semantic analysis.

4 Experimental Results

As mentioned in the introduction, in this paper we constructed a corpus consisted of 522 numbers of medical articles on Alzheimer disease published and written in Korean. For the purpose of performance evaluation on the modules for the NLP system developed in this paper, we selected a set of 78 counts of the domestic papers (among the corpora described above) as the test data set while the rest were taken as the training data set. Without the automated editors, for the most of the modules, we manually audited the test data set.

Using the test data set, we performed statistical analysis for the NLP modules suggested in the previous section. The values of the precision and the recall rates against the test data set are presented and summarized in the tables as below. Machine learning training error analysis is presented separately in Fig. 3.

In evaluation of the sentence detector, the test data set contains 140 sentences. The result of the test is presented in Table 1. In the table, the first row shows the precision and the recall rates of sentence detector, and the second row shows the success rates of the training.

Table 1: Evaluation of sentence detector

	precision	recall	data size
detector	95.71%	94.28%	140 sentences
trainer	96.85%	N/A	140 sentences

In evaluation of the tokenization module, the test data set contains 1,607 entities of the tokens. The result of the test is presented in Table 2 below. As seen in the table, the first row demonstrates the precision and the recall rates of tokenizer, and the second row demonstrates the success rates of the training.

Table 2: Evaluation of tokenization

	precision	recall	data size
detector	89.74%	92.96%	1,607 tokens
trainer	98.85%	N/A	1,607 tokens

In evaluation of the name finder module, the test data set contains 632 counts of the named entities. The result of the test is shown in Table 3 below. The precision and the recall rates are demonstrated in the first row, and the success rates of the training are demonstrated in the second row.

Table 3: Evaluation of name finder

	precision	recall	data size
classifier	100.00%	100.00%	632 N.E.
Trainer	99.92%	N/A	632 N.E.

In evaluation of POS tagger module, the test data set contains 1,607 counts of entities of the tags. The result of the tests is presented in the Table 4. The precision and the recall rates are shown at the first row, and the success rates of the training are in the second row.

Table 4: Evaluation of POS tagger

	precision	recall	data size
classifier	83.39%	79.71%	1,607 tags
trainer	87.86%	N/A	1,607 tags

In evaluation of chunker model, we sampled 54 counts of chunks from the test data set for NP, VP, and PP (preposition phrase), respectively. The results of the test are presented in Table 5. In the table, the precision and the recall rates estimated for the NP, VP, and PP are presented at the first, the second, and the row, respectively. At the fourth (last) row, the success rates of the trainer are presented. As seen in the third row, the number for the precision of PP is very low.

Table 5: Evaluation of chunker

	precision	recall	data size
NP	90.00%	94.74%	54 chunks
VP	75.00%	75.00%	54 chunks
PP	57.14%	100.00%	54 chunks
Trainer	86.59%	N/A	54 chunks

In evaluation of the parser, the test data set contains 154 phrases. The result of the test is presented in Table 6. In the table, the first row shows the precision and the recall rates of parser, and the second row shows the success rates of the training on the model.

Table 6: Evaluation of parser

	precision	recall	data size
classifier	72.58%	82.14%	154 phrases
trainer	94.94%	N/A	154 phrases

The modules consisting of NLP system presented in this paper are commonly based on the MAXENT by adopting error-correction rules [5, 6, 7, 8, 9, 10, 11]. Fig. 3 illustrates the convergence of iteration by error correction rules in terms of MSE (Mean Square Error) measure.

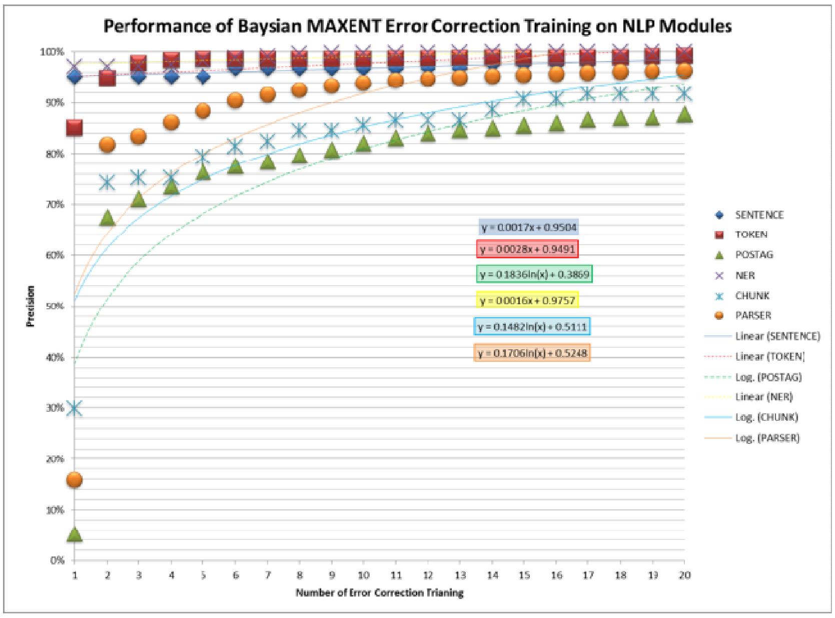

Fig. 3 Convergence of MAXENT based error correction used in the machine learning

As seen in the Fig. 3, except the name finder and the POS Tagger, the rest of the module reaches over 90% success rates. Our investigation showed that the low success rates of POS tagger are due to the over-determinedness of feature selection used in this study.

5 Conclusion

The purpose of this study is knowledge base system for the use in early diagnosis of Alzheimer disease. We presented in this paper the proposed design of NLP system and their implementations. Construction of corpus was also presented from the domestic paper on Alzheimer disease written in Korean.

The details are as follows: the first of all, we constructed the corpora of 522 counts of the published papers related with the Alzheimer disease. We implemented pdf-to-text conversion module in order to obtain text data automatically from pdf files in internet archive. Secondly, we developed a NLP system specialized in processing the medical literatures written in Korean language. The machine learning technique based NLP models are adopted from ASF OpenNLP. The NLP system presented in this paper is consisted of the 8 independent models: sentence detector, token detector, name finder, POS Tagger, word stem detector, chunker, parser, co-reference resolution model. Finally, test result showed that the most of NLP modules have high success rates of training (over 90%).

The test results show that NLP on the articles on Alzheimer disease performs as good as the general text data such as literatures.

For the future work, the NLP system presented in this paper will be used as a sentence-to-syntax conversion module in a large-scale automatic knowledge base system. As the well-known fact, natural language interpreter is the hardest part of the system as well as in HCI/AI system. To resolve the issues, we will continue our study on disambiguation in DRT, e.g., such as hole semantics. Our training data contains only 522 articles, we will continue working on construction of larger training data to end up with a well formed medical corpus written in Korean.

ACKNOWLEDGEMENTS

This research was supported by the Converging Research Center Program through the Ministry of Science, ICT and Future Planning, Korea (2013K000384).

References

[1] Sa-Kwang Song, Heung-Seon Oh, Sung-Hyon Myaeng, Sung-Pil Choi, Hong-Woo Chun, Yun-Soo Choi and Chang-Hoo Jeong. 2011. Procedural Knowledge Extraction on MEDLINE Abstracts, In proceedings of the 7th international conference on Active media technology(AMT 2011), 345-354.

[2] Kyung-Mi Park and Kyu-Baek Hwang. 2011. A Bio-Text Mining System Based on Natural Langauge Processing, Journal of KISS:Computing Practices and Letters, Vol.17, No.4, 205-213.

[3] B. Rosario and M. Hearst. 2004. Classifying semantic relations in bioscience texts, In proceedings of the annual meeting of the association for computational linguistics, 430-437.

[4] OpenNLP, http://opennlp.sourceforge.net

[5] Bishop C. 2006. Pattern Recognition and Machine Learning, Springer.

[6] David J. C. and MacKay. 2003. Information Theory, Inference, and Learning Algorithms, Cambridge University Press.

[7] E. Alpaydin. 2004. Introduction to Machine Learning, MIT Press: Cambridge.

[8] Lluis Marquez. 2000. Machine Learning and Natural Language Processing, Technical Report LSI-00-45-R.

[9] Richard O. Duda, Peter E. Hart and David G. Stork. 2000. Pattern Classification, Wiley-Interscience.

[10] Tom M. Mitchell. 2004. Machine Learning, McGraw Hill.

[11] A. Berger, S. Della Pietra and V. Della Pietra. 1996. A Maximum Entropy Approach to Natural Language Processing, Computational Linguistics, Vol.22, No.1.

[12] PennTags, http://tags.library.upenn.edu/

[13] P. Blackburn and J. Bos. 2005. Representation and Inference for Natural Language: a first course in computational semantics, CSLI publication, Stanford, CA.

[14] C.D. Manning, P.R. and H. Schütze. 2008. Introduction to Information Retrieval, Cambridge University Press.

[15] J. Gleik. 2012. The Information: A History, a Theory, a Flood, Vintage.

[16] E.A., Feigenbaum and P. McCorduck. 1983. The Fifth Generation (1st ed.), Reading, MA: Addison-Wesley.

[17] E. Charniak and E. Wilks. 1976. Computational Semantics. An Introduction to Artificial Intelligence and Natural Language Comprehension, vol. 4 of Fundamental Studies in Computer Science, North-Holland Publishing.

[18] H. Kamp and U. Reyle. 1993. From Discourse to the Lexicon: Introduction to ModelTheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory, Kluwer Academic Publisher.

[19] Deleger L1, Li Q, Lingren T, Kaiser M, Molnar K, Stoutenborough L, Kouril M, Marsolo K and Solti I. 2012. Building gold standard corpora for medical natural language processing tasks, AMIA Annu Symp Proc. 2012;2012:144-53, Epub.

[20] http://www-nlp.stanford.edu/links/statnlp.html

[21] A. Berger. 2001. Statistical machine learning for information retrieval, PhD Thesis, Carnegie Mellon University Computer Science Department Technical Report CMU-CS-01-110.

[22] J. Bos. 2011. A Survey of Computational Semantics: Representation, Inference and Knowledge in Wide-Coverage Text Understanding, Language and Linguistics Compass 5.6 (2011), 336-366.

[23] http://semanticweb.kaist.ac.kr/home/index.php/KoreanParser

[24] http://www.cs.utexas.edu/users/mfkb/RKF/km.html

[25] Refolo, L. M., Snyder, H., Liggins, C., Ryan, L., Silverberg, N., Petanceska, S. and Carrillo, M. C. 2012. Common Alzheimer’s Disease Research Ontology: National Institute on Aging and Alzheimer’s Association Collaborative Project, Alzheimer's & Dementia, Vol.8, No.4, 372-375.

[26] Krötzsch, M., Simancik, F. and Horrocks, I. 2012. A description logic primer, arXiv preprint arXiv:1201.4089.