Restoring an Elided Entry Word in a Sentence

for Encyclopedia QA System

Soojong Lim
Speech/Language
Information Research
Department
ETRI, Korea
isj@etri.re.kr

Changki Lee
Speech/Language Information Research Department
ETRI, Korea
leeck@etri.re.kr

Myoung-Gil Jang
Speech/Language Information Research Department
ETRI, Korea
mgjang@etri.re.kr

Abstract

This paper presents a hybrid model for restoring an elided entry word for encyclopedia QA system. In Korean encyclopedia, an entry word is frequently omitted in a sentence. If the QA system uses a sentence without an entry word, it cannot provide a right answer. For resolving this problem, we combine a rule-based approach with Maximum Entropy model to use the merit of each approach. A rule-based approach uses caseframes and sense classes. The result shows that combined approach gives a 20% increase over our baseline.

1 Introduction

Ellipsis is a linguistic phenomenon that people omit a word or phrase not to repeat a same word or phrase in a sentence or a document. Usually, ellipsis involves the use of clauses that are not syntactically complete sentences (Allen, 1995) but the fact does not apply to all cases. An ellipsis occurring in encyclopedia documents in Korean is an example.

(Entry word: Kim Daejung)

Korean: 공로[gongro]로[ro] 2000 년[nyeon] 노벨평화상 [nobel pyeonghwasang]을[eul] 받았다[batatda].

English: won the Nobel prize for peace in 2000 by meritorious deed.

In QA system(Kim et al, 2004), it answers a question using the predicate-argument relation as in the following example.

Korean: 2000 년[*nyeon*]에[*e*] 노벨평화상 [*nobelpyeonghwasang*] 을[*eul*] 받은 [*bateun*] 사람[*saram*]은[*eun*]?

English: Who’s the winner of the Nobel prize for peace on 2000?

받다(subj:사람, obj:노벨평화상, adv:2000 년)
(batda(subj:saram, obj: nobelpyeonghwasang, adv:ichunnyeon)
win(subj:who, obj:the Nobel prize for peace, adv:2000)

Entry word: 김대중
(Entry word: Kim Daejun)

받다(subj:NULL(김대중), obj:노벨평화상, adv:2000 년, 공로)
(batda(subj:NULL(kimdaejung), obj, nobelpyeonghwasang, adv: ichunnyeon, gongro)
win(subj:NULL(Kim Daejung), obj:the Nobel prize for peace, adv:2000, deed)

If an entry word of Korean encyclopedia performs a function of a subject or an objects, it is frequently omitted in the sentences of the Korean encyclopedia. If the QA system uses the result in the above example, it cannot find who won the Nobel prize for peace in the year of 2000. We need to restore an entry word as a subject or an object to answer a right question.

In this paper, to overcome this problem, we first try to classify entry words in encyclopedia into sense classes and determine which sense classes are restored to the subjects or the objects. Then we use caseframes for determining sense classes which are not restored using sense classes. If there is no caseframes, we use a statistical method, ME model, for determining whether the entry word is restored or not. Because each approach has both strength and weakness, we combine three approaches to achieve a better performance.

2 Related Work

Ellipsis is a pervasive phenomenon in natural languages. While previous work provides important insight into the abstract syntactic and semantic representations that underlie ellipsis phenomena, there has been little empirically oriented work on ellipsis.

There are only two similar empirical experiments done for this task. First is Hardt's algorithm(Hardt, 1997) for detecting VPE in the Penn Treebank. It achieves precision levels of 44% and recall of 53%, giving an F-Measure of 48% using a simple search technique, which relies on the annotation having identified empty expressions correctly. Second is Nielsen's machine learning techniques(Nielsen, 2003). They only try to detect of elliptical verbs using four different machine learning techniques, Transformation-based learning, Maximum entropy modeling, Decision Tree Learning, Memory Based Learning. It achieves precision levels of 85.14% and recall of 69.63%, giving an F-Measure of 76.61%. There are 4 steps: detection, identification of antecedents, difficult antecedents, resolving antecedents. Because this study only concentrates on the detection, a comparison with our study is inadequate.

We combine rule-based techniques with machine learning technique for using the merit of each technique.

3 Restoring an Elided Entry Word

We use three kinds of algorithms: A caseframe algorithm, an acceptable sense class algorithm, and Maximum Entropy (ME) algorithm. For knowing a strength and weakness points of each algorithm, we do experiments on each algorithm. Then we combine algorithms for higher performance.

Our system answers in three ways: restoring an entry word as a subject, restoring an entry word as an object, and does not restore an entry word. We evaluate an algorithm in two ways. First, we evaluate all answers with precision. Second, we evaluate just two answers, restoring an entry word as a subject and object, with F-measure.

$$recall = \frac{correct\ elided\ entry\ words\ found}{all\ elided\ entry\ words\ in\ test\ set}$$

$$precision = \frac{correct\ elided\ entry\ words\ found}{all\ elided\ entry\ words\ found}$$

$$F - measure = \frac{2 \times precision \times recall}{precision + recall}$$

3.1 Using Caseframes

We use modified caseframes constructed for Korean-Chinese machine translation. The format of Korean-Chinese machine translation case frame is as the following:

A=Sense_code!case_particle verb > Chinese > Korean Sentence

A=사람(*saram*)!가(*ga*) B=장소(jangso)!로(ro) 가(ga)!다(da) > A 0x53bb:v B [그(geu)[A]가(ga) 바다(bada)[B]로(ro) 가다(gada)]

A=Person!subj B=Location!adv go.

In the caseframe, we only use Sense Class, case particle marker, and the verb. The caseframe used in this research consists of 30,000 verbs and 153,000 caseframes.

The sense class used in this research is selected from the nodes of the ETRI Lexical Concept Network for Korean Nouns which consists of about 60,000 nodes. (If we include proper nouns, the total entry of ETRI Lexical Concept Network for Korean Nouns is about 300,000 nodes).

First, we analyze a sentence using dependency parser (LIM, 2004), and then we convert a result of a parser into the caseframe format. We determine to restore an entry word if there is an exactly matched caseframe of a target except a sense class of an entry word.

Table 1 shows an example.

First, we analyze a sentence using dependency parser (LIM, 2004), and then we convert a result of a parser into the caseframe format. We determine to restore an entry word if there is an exactly matched caseframe of a target except a sense class of an entry word.

Table 1. An Example of Caserframe Algorithm

Input	Entry word: Along Bay Sense: Location Sentence: Located in East of Haiphong
Parsing	Locate(subj:NULL, obj:NULL, adv: east of Haiphong) Caseframe of sentence direction!*e* locate
Matching	24265-2 A=Location!ga B=Location!eseo C=*direction*!*e* 24265-4 A=Location!ga B=*direction*!e 24265-8 A=weather!ga B=*direction*!*e* 24265-12 A=*direction*!*e* 24265-17 A=body!ga B=*direction*!*e*
decision	Restoring an entry word as a subject

The result of caseframe algorithm is in table 2. The result of caseframe algorithm shows that it has a high precision but a relatively low recall because it is impossible to construct caseframes for all sentences.

Table 2. Result of Caseframe Algorithm

	Subject	Object	Sum
Precision	88.16	6.38	56.91
Recall	59.29	27.28	56.45
F-measure	70.90	10.34	56.68

3.2 Acceptable Sense Class

All entry words in the encyclopedia belong to at least one sense class. We verify all 444 sense classes to see whether they could be restored in a sentence. We set a precision threshold 50% and we fix 36 sense classes to “acceptable sense class”. An acceptable sense class is a sense class that if an entry word is included in an acceptable sense class, we unconditionally restore an entry word in a sentence. Our verification tells that there is only acceptable sense classes for subjects. Table 3 shows acceptable sense classes.

Table 3. Acceptable Sense Classes

PERSON, ORGANIZATION, STUDY, WORK, LOCATION, ANIMAL, PLANT, ART, BUILDING, BUSINESS MATTERS, POSITION, SPORTS, CLOTHES, ESTABLISHMENT, PUBLICATION, MEANS of TRANSPORTATION, EQUIPMENT, SITUATION, HARDWARE, BROADCASTING, HUMAN RACE, EXISTENCE, BRANCH, MATERIAL OBJECT, WEAPON, EXPLOSIVE, LANGUAGE, FACILITIES, ACTION, SYMBOL, TOPOGRAPHY, ROAD, ECONOMY, ADVERTISEMENT, EVENT, TOMB

The result of acceptable sense class algorithm is presented in table 4. Because we cannot get acceptable sense classes for objects, F-measure of object is 0.

Table 4. Result of ASC Algorithm

	Subject	Object	Sum
Precision	58.14	0.0	58.14
Recall	66.37	0.0	60.48
F-measure	61.98	0.0	59.29

3.3 Maximum Entropy Modeling

Maximum entropy modeling uses features, which can be complex, to provide a statistical model of the observed data which has the highest possible entropy, such that no assumptions about the data are made.

$$p^* = \arg\max_{p \in C} H(p)$$

where p^* is the most uniform distribution, C is a set of probability distributions under the constraints and $H(p)$ is entropy of p.

Ratnaparkhi(Ratnaparkhi 98) makes a strong argument for the use of maximum entropy modes, and demonstrates their use in a variety of NLP tasks.

The Maximum Entropy Toolkit was used for the experiments.[1]

Because maximum entropy allows for a wide range of features, we can use various features, such as lexical feature, POS feature, sense feature, and syntactic feature. Each feature consists of subfeatures:

Lexical feature;
Verb_lex : lexeme of a target verb
Verb_e_lex : lexeme of a suffix attatched to a target verb

POS feature;
Verb_pos : pos of a target verb
Verb_e_pos : pos of a suffix attatch to a target verb

Sense feature;

[1] Downloadable from http://homepages.inf.ed.ac.uk/s0450736/maxent_toolkit.html

Ti_res_code: where sense of an entry word is included in acceptable sense class
Verb_cf_subj, obj: whether a sense of entry word is included in caseframe of a targe verb
Ti_sense : sense class of entry word

Syntactic feature;
Tree_posi: position of parse tree
Rel_type: relation type between verbs in a sentence
Sen_subj, sen_obj : existence of subject or object

Hybrid feature;
Pair =(sense class of entry word, verb)

Table 5 shows an example of features that we use for finding an elided entry word.

Previous work using ME model adopted distance-based context for training. Because we use syntactic features, we can use not only distance-based context but also predicate-argument based context. The training data for ME algorithm consist of verbs in the encyclopedia document and their syntactic arguments. Each verb-arguments set is augmented with the information that signifies whether a subject, an object or neither of them should be restored. For training, we use a dependency parser[Lim, 2004]. A precision of this parser is about 75%. The results of ME model algorithm is shown in table 6. The results of ME model shows that its score is the lowest of all. We guess the reason is that there is not enough training data for covering all sense classes.

Table 5. An Example of Features

Entry word, Sentence	!TI Cirsotrema perplexam !SENSE Animal !VERB live !SENT lives in a tidal zone
Lexical feature	verb_lex=살다(*salda*) verb_e_lex=며 (*myeo*)
POS feature	verb_pos=4 verb_e_pos=24
Sense feature	ti_res_code=1 verb_cf_subj=1 verb_cf_obj=0 ti_sense=Animal
Syntactic feature	tree_posi=high rel_type=-1 sen_subj=0 sen_obj=0
Hybrid feature	pair=(Animal, live)

Table 6. Result of ME Model

	Subject	Object	Sum
Precision	62.50	40.0	60.87
Recall	35.40	18.18	33.87
F-measure	45.20	25.00	43.52

3.4 Combining Algorithms

Different algorithms have different characteristics. For example, the acceptable sense class algorithm has relatively high recall but low precision, while the opposite holds true for the caseframe algorithm, we need to combine algorithms for maximizing advantages of each algorithm.

First, we combine the acceptable sense class algorithm with the ME model. We process the problem using the sense class algorithm. Instead of applying the algorithm exactly, we use the ME model for helping the acceptable sense class algorithm. If the acceptable sense class algorithm determines a restoration, we adopt the case to ME model. Then if the score of ME model is over the negative threshold, we determine not to restore an entry word.

Second, we combine the caseframe algorithm with the ME model. We process the cases not resolved in the first processing time using the caseframe algorithm. We try to match caseframes exactly to sentence with an entry word sense code. If we cannot find the exactly matching caseframe, we try matching caseframes partially. In this case, a precision is maybe lower than an exact match, we also use the ME model for reliability. If the score of ME model is over the positive threshold, we determine to restore an entry word.

4 Result and Conclusion

For ME model, we made a training set manually. The training set consists of 2895 sentences: 916 sentences for restoring an entry word as a subject, 232 sentences for restoring an entry word as an object, 1756 sentences for not restoring any. For a test, we randomly selected 277 sentences.

We did 6 kinds of experiments. Using Caseframe algorithm(CF), Acceptable sense class algorithm(ASC), ME model(ME) and combine ASC with CF(ASC_CF), ASC with ME

(ASC_ME), and ASC with CF and ME(ASC_CF_ME).

Table 7. Result of Combined Algorithm

	Recall	Precision	F-measure
baseline	100.00	31.64	48.07
ASC_CF_ME	78.23	60.25	68.07
ASC_CF	68.55	50.00	57.82
ASC_ ME	79.03	59.39	67.82

The performance of the methods is calculated using recall, precision and F-measure.

Table 7 and Figure 1 show the performance of each experiment.

Our proposed approach (ASC_CF_ME) gives the best results among all experiments, with an F-measure of 68.1%, followed closely by ASC_ME. This gives a 20% increase over our baseline. For testing a portability of our approach, we experiment the noun phrase ellipsis (NPE) detection. The performance of NPE is alike an elided entry word. Recall is 69.31, Precision is 65.05, and F-measure is 67.12. So we expect the performance of our approach not to drop when applied to NPE or other ellipsis problem. The results so far are encouraging, and show that the approach taken is capable of producing a robust and accurate system.

In this paper, we suggested the approach that restores an elided entry word for Encyclopedia QA systems combining an acceptable sense class algorithm, a caseframe algorithm, and ME model.

For future work, we plan to pursue the following research. First, we will use various machine learning methods and compare them with the ME model. Second, because we plan to apply this approach in the encyclopedia document, we need to design the more general approach to use other ellipsis phenomenon. Third, we try to find a method for enhancing performance of restoring elided entry words as the object.

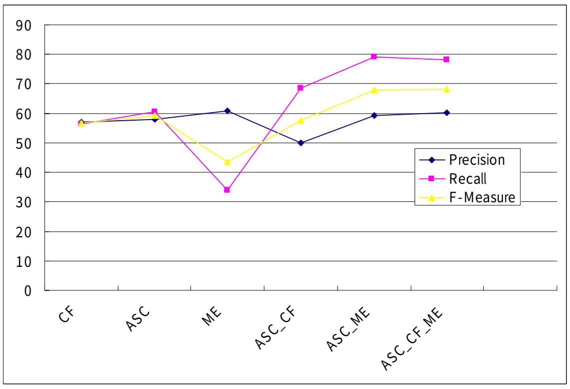

Figure 1. Comparison of All Results

References

James Allen. 1995. *Natural Language Understanding,* Benjamin/Cummings Publishing Company, 449~455

Leif Arda Nielsen. 2003. *Using Machine Learning Techniques for VPE detection*, RANLP 03, Bulgaria.

Daniel Hardt. 1997. *An empirical approach to vp ellipsis*, Computational Linguistics, 23(4).

Adwait Ratnaparkhi. 1998. *Maximum Entropy Models for Natural LANGUAGE Ambiguity Resolution*, Unpublished PhDthesis, University of Pennsylvania.

Lim soojong. 2004. *Dependency Relation Analysis Using Caseframe for Encyclopedia Question-Answering Systems*, IECON, Korea.

H. J. Kim, H. J. Oh, C. H. Lee., et al. 2004. *The 3-step Answer Processing Method for Encyclopedia Question-Answering System: AnyQuestion 1.0.* The Proceedings of Asia Information Retrieval Symposium (AIRS) 309-312