Advanced Science and Technology Letters
Vol.120 (GST 2015), pp.467-471
http://dx.doi.org/10.14257/astl.2015.120.91

A Multi-pass Coreference Resolution Model using Context Restriction

Mei-ying Ren[1], Sinjae Kang[1]

Dept. of Computer & Information Engineering, Daegu University
201, Daegudae-ro, Gyeongsan-si, Gyeongsangbuk-do, 38453, Republic of Korea
{meeyeong1211@hotmail.com, sjkang@daegu.ac.kr}

Abstract. In this paper, we present a multi-pass coreference resolution model using context restriction. Coreference resolution is the task of finding all expressions such as words or phrases that refer to the same entity in a text. First, we classified coreference types into seven according to the accuracy of the types, and then modelled a sequential rule-based approach that restricts global or local context ranges at each step. Since anaphors generally refer to the nearest preceding individuals with same attributes, position information is one of the important features in coreference resolution.

Keywords: Coreference Resolution, Context Restriction, Global Context, Local Context, Rule-based System.

1 Introduction

Coreference is a relationship between two words or phrases in which both refer to the same real-world entity, and serves an important role of linking related information together. It consists of two linguistic expressions - antecedent and anaphor, which are called mentions. Coreference resolution is a task to discover the antecedent for each anaphor in a text. The task is important for high-level NLP (Natural Language Processing) that involves natural language understanding such as machine translation, document summarization, and QA (question & answering).

Recent researches on coreference resolution have used lexical, syntactic features and global inference which performing coreference resolution for all mentions in a text. The research showing state-of-the-art performance on coreference resolution in English is proposed by Lee, et al [1]. The paper analyzed different cases of the issue and defined various sieves. Then they proposed to start from high precision sieve to get both high precision and recall. [2] applied the multi-pass sieve model in Korean and also obtained quite higher performance. There is other research on Korean used SVM and mention-pair method [3], but didn't show better results than [2].

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (NRF-2011-0007025).

ISSN: 2287-1233 ASTL

In coreference resolution, position information plays an important role. According to [3, 4, 5], position information was one of the significant features of the mention. It is due to that the anaphor commonly refer to the closest antecedent that has the same attributes.

In this paper, we proposed a multi-pass model that applies different context ranges at each pass. It is determined based on the characteristics of each coreference type.

2 A Proposed Coreference Resolution Model

We re-defined Korean coreference resolution issue. In order to reduce errors existed in the previous studies, we applied different context ranges at each pass, such as paragraph level (local context) or passage level (global context). Figure 1 shows the proposed model of multi-pass Korean coreference resolution.

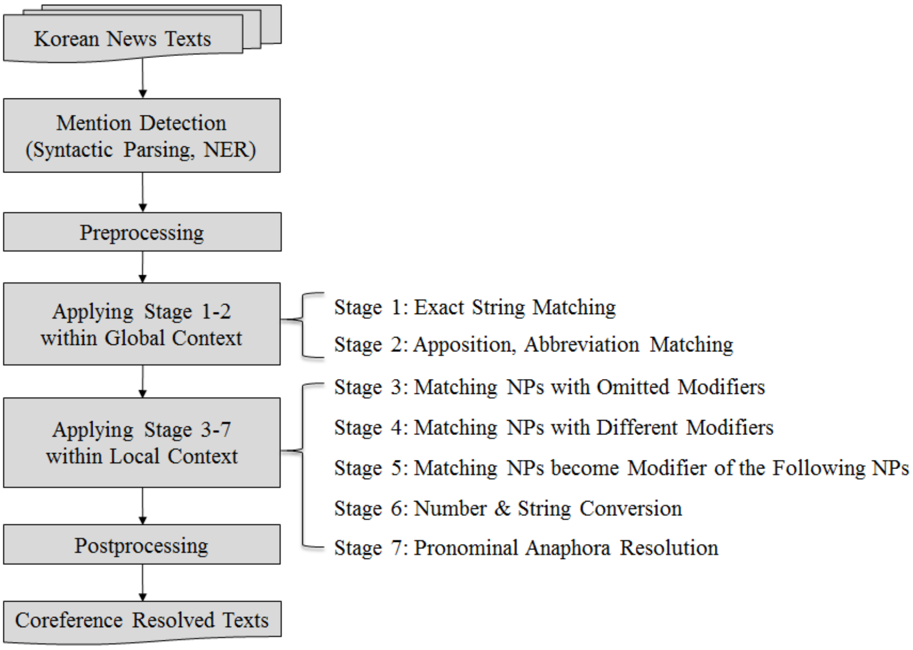
Korean News Texts
Mention Detection (Syntactic Parsing, NER)
Preprocessing
Applying Stage 1-2 within Global Context
Applying Stage 3-7 within Local Context
Postprocessing
Coreference Resolved Texts
Stage 1: Exact String Matching
Stage 2: Apposition, Abbreviation Matching
Stage 3: Matching NPs with Omitted Modifiers
Stage 4: Matching NPs with Different Modifiers
Stage 5: Matching NPs become Modifier of the Following NPs
Stage 6: Number & String Conversion
Stage 7: Pronominal Anaphora Resolution

Fig. 1. Proposed method for Multi-pass Korean Creference Resolution

In the first mention detection step, noun phrases are extracted from Korean news texts which are syntactically analyzed and segmented by paraphrase units. In preprocessing, four simple processes are applied to filter unnecessary noun phrases. The first process uses a longest matching technique from the rear of mentions to deal with the mentions whose last words are in the same position. The second process extracts NERs and attaches tags. The third process is pronoun designation. The last process eliminates mentions contain stopwords such as “등 (*dung/such as*)”, “수

(*su/able to*)", "것 (*gut/the postposition converting verbs to nouns*)", and date expressions.

For the stage 1-2, we apply the matching rules over entire passage, i.e. global context. Stage 1 is matching exactly same strings. Stage 2 is matching appositions and abbreviations. Apposition is the case that the name and title or job appears together. For example, " 김주형 (*Kim Joo-hyeong*)" and " 김주형 박사 (*Dr. Kim Joo-hyeong*)" would be matched at this stage. In Korean, there are various abbreviations. We classified abbreviations into four types —" noun combination" type, " syllable combination" type, " noun combination and syllable combination mix" type and "personal abbreviation" type. Moreover, since English acronyms also appear in Korean text, these are also treated in Stage 2.

The exactly matched mentions can be assumed as the most precise cases, so we conduct the Stage 1 within passage level. In Stage 2, appositions and abbreviations can refer a subject clearly with NER information, so these are also qualified to be conducted within passage level.

The remaining 3-7 stages are conducted within paragraph level, i.e. local context. Table 1 listed the conditions and examples of the remained 3-7 stages should be conducted within local context level.

Finally, in postprocessing, we removed mentions not involved in the previous coreference resolution stages.

Table 1. Coreference Types Resolved in Local Context

Stage	Process	Condition	Example
3	Matching NPs with Omitted Modifiers	MOD +NP NP MOD+NP must appear in front of NP	나이지리아 남부 아바 마을 (*Nigeria nambu aba ma-ul /Aba village located in south of Nigeria*) = 마을에서는 (*ma-ul-e-so-nun/In the village*) 얼어붙은 연극계 (*el-o-bu-tun youn-guk-gye/the frozen theatrical fields*) = 연극계 (*youn-guk-gye/the theatrical fields*)
4	Matching NPs with Different Modifiers	MOD1 + NP MOD2 + NP MOD2 should include MOD1 (String Level)	최저 생계비 (*choi-je saeng-gye-bi/ the minimum cost of living*) = 50만원의 최저 생계비 (*o-shib manwon-eui choi-je saeng-gye-bi/ ₩500,000 of the minimum cost of living*) 예술단 구조조정 (*ye-sul-dan gu-jo-jo-jeng/restructuring in the art centers*) = 3개 예술단의 구조조정 (*se-gae*

			ye-sul-dan-eui gu-jo-jo-jeng/ restructuring in the 3 art centers)
5	Matching NP becomes Modifier of the Following NP	NP1 NP1 +NP2 No postpositional particles between NP1 and NP2 One of NPs should be NER	같은 기획사 (*gatun gihoeksa/ the same entertainment company*) = 같은 기획사 SM (*gatun gihoeksa SM/ the same entertainment company SM*)
6	Number& String Conversion	Same NP Number should be above 2.	세 단체 (*se-dan-che/ three associations*) = 3개 예술단 단체 (*se-gae ye-sul-dan dance/ 3 art associations*) = 3개 공연 단체 (*se-gae gong-youn-danche/ 3 performance associations*)
7	Pronominal Anaphora Resolution	Personal pronoun	김주형 (*Kim Joo-hyung*) = 김주형 박사 (*Kim Joo-hyung baksa/ Dr. Kim Joo-hyung*) = 그 (*gu/ he*)

3 Conclusion

In order to overcome the problem caused by applying global inference which performing coreference resolution for all mentions in a text, this paper proposed a multi-pass coreference resolution model that applies different context ranges at each stage. According to the characteristics of each coreference type, the 1-2 stages are applied within passage level (global context), and the 3-7 stages applied within paragraph level (local context) including text titles.

References

1. Lee, H., Chang, A., Peirsman, Y., Chambers, N., Surdeanu, M., Jurafsky, D.: Deterministic Coreference Resolution Based on Entity-Centric, Precision-Ranked Rules. Computational Linguistics, vol. 39 (4), 885--916 (2013)
2. Park, C., Choi, K., Lee, C.: Korean Coreference Resolution using the Multi-pass Sieve. Journal of KIISE, vol. 41 (11), pp. 992-1005. (2014)
3. Park, C., Choi, K., Lee, C.: Coreference Resolution for Korean using Mention Pair with SVM. KIISE Transactions on Computing Practices, vol. 21 (4), pp. 333-337. (2015)
4. Bengtson, E., Roth, D.: Understanding The Value of Features for Coreference Resolution. Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 294-303. Association for Computational Linguistics (2008)

5. Luo, X., Ittycheriah, A., Jing, H., Kambhatla, N., Roukos, S.: A Mention-Synchronous Coreference Resolution Algorithm Based on the Bell Tree. Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, p. 135. Association for Computational Linguistics (2004)
6. Yoon, Y., Song, Y., Lee, J., Lim, H.: Construction of Korean Acronym Dictionary by Considering Ways of Making Acronym from Definition. Proceedings of Spring Conference for KSCS, pp 81-85. The Korean Society for Cognitive Science (2006)