Journal of Computer Science

Original Research Paper

Anaphora Resolution in Thai EDU Segmentation

Authapon Kongwan, Siti Sakira Kamaruddin and Farzana Kabir Ahmad

School of Computing, Universiti Utara Malaysia, Malaysia

Article history
Received: 30-12-2021
Revised: 07-03-2022
Accepted: 31-03-2022

Corresponding Author:
Siti Sakira Kamaruddin
School of Computing,
Universiti Utara Malaysia,
Malaysia
Email: sakira@uum.edu.my

Abstract: Human knowledge is mostly in the form of unstructured text. Text can be transcribed into various languages such as the Thai language. To extract knowledge from Thai text, natural language tasks such as word segmentation, Elementary Discourse Unit (EDU) segmentation, and anaphora resolution is the needed tasks. Some interesting phenomena such as non-referential anaphora and the ellipsis of the owner are the significant problems that are necessary to resolve before constructing the complete semantic in the Natural Language Processing (NLP) application. The non-referential anaphora must be detected before identifying the referential anaphora to improve the precision of the anaphora resolution. The ellipsis of the owner is also a crucial problem that needs to be resolved to find the complete semantics. This study presents the methodology to resolve the anaphora from Thai EDU segmentation. The methodology is divided into 2 parts: Thai morphological analysis and the anaphora resolution. The ranking model is applied to resolve the reference of anaphora with the features from the surface word, surround word, syntactic information, and ontology. The results show that precision is 0.77, recall is 0.84 and the F1 score is 0.81.

Keywords: Anaphora Resolution, Thai Anaphora, Ranking Model, Natural Language Processing

Introduction

Text is a significant source of human knowledge. Most human knowledge is in the form of unstructured text. The research areas that are concerned with knowledge identification in text such as Information Extraction (IE), Knowledge Extraction (KE), and Question Answering System (QAS) need Natural Language Processing (NLP) to identify the interesting pieces of the information to construct a knowledge-based. Natural Language Processing tasks are a crucial part to achieve that goal, especially in Thai text (Netisopakul and Wohlgenannt, 2017; 2018).

Thai text processing is a challenging task to achieve. The Thai word boundary identification is the first challenging task to be completed. The Thai text can be looked like a stream of continuous characters in a paragraph without any space character or punctuation. There are some features such as the absence of words and unclear word boundaries that make this task more complicated to process (Aroonmanakun, 2007). The Thai word segmentation (Kongyoung *et al.*, 2015; Boonkwan and Supnithi, 2017) is still active research in the Thai text processing task.

The Thai sentence boundary identification (Slayden *et al.*, 2010; Zhou *et al.*, 2016) is also a non-trivial task in Thai NLP tasks. The Thai sentence can be written continuously in a paragraph without space or explicit marker to indicate the sentence boundary. To construct the knowledge from text, the sentence in Thai text needs to be processed to indicate the boundary and then specify the semantic concept and build the semantic relation. However, in some applications such as text summarization (Sukvaree *et al.*, 2007; Ketui *et al.*, 2015), the smaller unit, which is called an Elementary Discourse Unit (EDU) (Marcu, 1998, 1999; Carlson *et al.*, 2003), can be more suitable to process rather than the sentence. Thai EDU segmentation research (Ketui *et al.*, 2013; Kongwan *et al.*, 2020) is still in progress to gain higher precision in identifying the EDU segment.

Anaphora resolution is an NLP task that solves the referent objects in text. The anaphora resolution research in Thai text is still rare (Aroonmanakun, 2000; Pathanasin, 2018). To find the complete semantics in Thai text, the anaphora resolution with acceptable precision is an essential key to success. Some phenomena are interesting problems that appear in Thai text on anaphora resolution. There are two crucial problems that we would mention the non-referential anaphora and the ellipsis of the owner. In the text, some anaphoras do not refer to any object but refer to the reader or the generalized object. The anaphora that do not refer to any object in the text is called non-referential anaphora. The non-referential anaphora must be detected before identifying the referential anaphora to improve the precision of the semantic structure. Then, before resolving the reference of the anaphora, we need to identify whether the anaphora is a non-referential anaphora or not. Moreover, some parts of the object can be omitted in Thai text such as the preposition of the owner. The omission of the preposition of the owner is called the ellipsis of the owner. The ellipsis of the owner is the language phenomenon that needs to be resolved to get complete information from the text. Due to the complicated sentence breaking, the anaphora resolution in the EDU segmentation can be more useful. Discourse relation is the relationship between the discourse segment. Discourse relation is needed in the NLP application such as text summarization. However, it is possible to resolve the reference of the anaphora by not using the discourse relation. This study will experiment with resolving the anaphora on Thai EDU segmentation with no discourse relation involved.

Anaphora and Coreference Resolution

This section presents the summary of the methodology collected from some good review papers (Poesio *et al*., 2016; Sukthanker *et al*., 2020). The methodology of anaphora and coreference resolution can be categorized into rule-based and learning-based as follows.

Rule-Based

Rule-based anaphora resolution is based on hand-crafted rules. The rules are based on syntactic and semantic features that are related to the text. Hobb's algorithm (Hobbs, 1978) is the proposed algorithm to resolve pronouns with rules on the syntactic parse tree.

The algorithm traversal on the syntactic parse tree of the sentence with a breadth-first search for an antecedent and prune the antecedent search space with rules and selection constraints. Lappin and Leass's algorithm (Lappin and Leass, 1994) is a knowledge-rich algorithm that incorporates the theories of salience. The candidates are filtered by using the syntactic information with binding constraints and then calculating the salience weight. The candidate with the highest salience weight is selected to determine the result. Although most of the rule-based approaches are rich in knowledge, there is some research (Lee *et al*., 2013; Zeldes and Zhang, 2016) that intends to work on reducing the dependency of the rule on external knowledge.

The centering theory (Grosz *et al*., 1995) is an algorithm that interprets phenomena like anaphora and coreference in the discourse structure in terms of centers. Centers are discourse entities that are referred to as utterances in the discourse segment. The forward-looking Centers (Cf) are a set of centers that are realized in the utterance. The backward-looking Center (Cb) is referred to as a center of attention belonging to the set of the Cf in the current and the preceding utterances. The algorithm starts by finding all of the possible discourse entities in utterances as the Cf. One of the Cf would define the Cb of the utterance by the highest rank that is realized from some constraints and rules. The centering theory can be used not only in English. The other languages such as Japanese (Iida, 1996), Italian (Di Eugenio, 1998), German (Strube and Hahn, 1996), and Thai (Aroonmanakun, 2000) can also use the center theory by using the same constraints and rule with some modifications.

Building comprehensive rules in the rule-based anaphora resolution is difficult because those rules are based on hand-craft building. The corpus changing may affect the rules that were built from the prior corpus. The learning-based solution could be easier to produce comprehensive rules on the corpus changing.

Learning-Based

The learning-based approach to anaphora and coreference resolution come to an impact in the late nineties. The learning-based such as decision trees (Aone and William, 1995), genetic algorithms (Mitkov *et al*., 2002), and Bayesian rule (Ge *et al*., 1998) is the early algorithms that are used to resolve the anaphora resolution. The learning-based models on anaphora and coreference can be classified into four groups that are mention-pair, entity-mention, ranking model, and deep learning model.

The coreference in the mention-pair model is organized as a collection of NP's pair links. The model uses a classification to deal with the pair links to find which pair is a reference. Decision trees and random forests (Lee *et al*., 2017a) are widely implemented as classifiers for anaphora and coreference resolution. Also, the statistical learners (Ge *et al*., 1998), memory learners (Daelemans *et al*., 2004), and rule-based learners (Cohen and Singer, 1999) are also popularly implemented. The mention-pair model also works on generating an NP partition for coreference chains. Clustering techniques are implemented for this task such as best-first clustering (Ng and Cardie, 2002), closest-first clustering (Soon *et al*., 2001), correlational clustering (McCallum and Wellner, 2004), Bell Tree beam search (Luo, 2005) and graph partitioning algorithms (Nicolae and Nicolae, 2006).

The entity-mention model utilizes the prior coreference decision to link with a target entity instead of an antecedent. The classifier is modified to learn whether the pair of NP assigned to a partial cluster is positive or negative. There is a comparison of entity-mention and mention-pair models that uses the decision trees and inductive logic programming. The results of the entity-mention model are not better than the mention-pair model. The major problem is that it is very difficult to define the features on the cluster for the entity-mention model. There are recent works (Clark and Manning, 2016b; Liu *et al*., 2020) that attempt at learning cluster-level features for the entity-mention model.

DOI: 10.3844/jcssp.2022.306.315

The prior models are working on the binary classifier that decides whether an antecedent is a coreference or not. The ranking model is working on ranking the mention and then choosing the best candidate to be a coreference. This algorithm is a more natural way to determine the coreference between the different antecedents. There are notable works (Denis and Baldridge, 2008; Durrett and Klein, 2013) that work on the ranking model by changing the binary classifier to the ranking model.

The deep learning model is also a new method to reduce the dependency on hand-craft features in coreference resolution. Words are represented as vectors conducting the semantic dependencies (Pennington *et al*., 2014). Techniques in mention-pair, entity-mention, and ranking models are adapted to training in the neural network. Clark and Manning (2016a) and Lee *et al*. (2017b) 's works that have done with these techniques.

The learning-based approach to anaphora and coreference resolution gives a good result and be easier to change the corpus or domain. The same model can be adapted to a new language easily with the minor adaptation of the feature sets. The ranking model, which resolves the anaphora by choosing the best candidate from the antecedents, produces higher precision results compare to the other models.

Anaphora in Thai Texts

The anaphora is a linguistic tool for referencing a thing mentioned earlier in a discourse. A phenomenon like non-referential anaphora is an interesting item that affects the anaphora resolution in this study. The interesting information on the use of anaphora in Thai text is described in this section.

Anaphora Types

In this study, we define the anaphora in 4 types which are zero anaphora, pronominal anaphora, nominal anaphora, and ellipsis of the owner. All types of anaphora are described as follows.

Zero Anaphora

Zero anaphora is the use of a gap in the subject of a sentence that references the object in the prior sentence. There is normally a lot of use of zero anaphora in Thai text. Due to the use of zero anaphora, a Thai sentence can be formed by only a verb phrase. In the process of EDU segmentation, the embedded relative clause EDU can form a zero anaphora after EDU segmentation.

Pronominal Anaphora

Pronominal anaphora is the use of pronouns to refer to the object in the prior sentence. A pronoun is a fundamental linguistic tool to refer to the thing that has been introduced in the antecedent. The use of the pronoun is widely used in the corpus. The resolution of the pronoun may need additional information such as gender, and number to resolve the reference.

Nominal Anaphora

Nominal anaphora is the use of nouns with a determiner to refer to the object in the prior sentence. A noun that is nominal anaphora can be a supertype (hyponymy) of the reference. A determiner can be used as an indication to identify the nominal anaphora. This anaphora can be resolved with the utilization of the semantic ontology to resolve the hyponymy.

Ellipsis of the Owner

Nouns in Thai text can omit the preposition of the owner that was introduced in the antecedent. Mostly, a part-of or meronymy is a semantic relation that attaches between a noun and the ellipsis. Additional information like the ontology of meronymy is needed for resolving the ellipsis of the owner.

Referential and Non-Referential Anaphora

Anaphora generally refer to the reference object in the antecedent. There is an interesting phenomenon that the anaphora may not refer to any object in text. Therefore, the anaphora can be tagged into 2 kinds that are referential and non-referential anaphora.

Referential Anaphora

Referential anaphora means any type of anaphora that refers to the object in the text. Mostly, the anaphora that appear in the text is the referential anaphora. From the observation in the corpus, the pronoun, zero anaphora, and ellipsis of the owner mostly refer to the existing entities in the text. However, there is a lot of nominal anaphora that do not refer to any object in the text. Before resolving the referential anaphora, the anaphora should be identified whether it is referential or non-referential anaphora.

Non-Referential Anaphora

Non-referential anaphora means any type of anaphora that does not refer to any explicit entity in text. Any type of anaphora can be a non-referential anaphora. In zero anaphora, non-referential anaphora occurs mostly from the use of the verb of occurrence. Some verbs can generate the non-referential anaphora in zero anaphora such as ”เกิด(occur, birth)”, ”มี(happen, has)” and ”เป็น(be)”. There is the pronoun ”เรา(we)" that can refer to the reader or general people that does not refer to any object in the text. In nominal anaphora, there is the word with some determiner that refers to the general object that is not specified to any object in the text. The surface word could be used for learning to identify which nominal anaphora could be non-referential.

Methodology

The implementation of training and resolution in all parts of this study is implemented by Golang to ensure high performance and memory usage efficiency. All the processes are computed on a computer server with Intel(R) Xeon(R) Silver 4114 CPU @ 2.20GHz 16GB memory. The methodology is divided into 2 parts. The first is the Thai morphological analysis for data preparation and the second is the anaphora resolution.

Thai Morphological Analysis

In this study, Thai morphological analysis is processed from Thai word segmentation to Thai EDU segmentation following the process from the previous work (Kongwan *et al*., 2020). The data source is from the Thai Wikipedia webpage. The selected pages are downloaded to store in a database and then pass through the corpus cleaning process to remove the HTML tag and some unused information in pages. After that, some symbols in pages were converted to symbol tags to produce the cleaned corpus. After the cleaning process, the corpus is submitted to process the Thai word segmentation, Thai named entities identification, and then Thai EDU segmentation.

Anaphora Resolution

There are 3 steps of processes in the anaphora resolution: Anaphora determiner, resolution for non-referential anaphora, and resolution for referential anaphora. Anaphora determiner is the algorithm for determining the anaphora type in EDU. After that, the resolution for non-referential anaphora is applied to distinguish the anaphora which is the non-referential or referential anaphora. Finally, the resolution for referential anaphora is applied to find the reference of the referential anaphora from the antecedent EDU. The ontology is a background knowledge that contains semantic concepts and semantic relations such as meronymy and hyponymy. The ontology is significant in the anaphora determiner and is a component of the feature set for the anaphora resolution process. Figure 1 shows the overview of the anaphora resolution processes.

Corpus Preparation

The corpus for anaphora resolution has come from the result of The EDU segmentation process. The corpus will be tagged with the additional information for training in the anaphora resolution training model. The entities in the corpus will be tagged with the number for reference. Each anaphora will be tagged with the number and the reference number. A zero will be tagged in the reference number in the case of the non-referential anaphora. Figure 2 shows the example of the anaphora tagging in the corpus.

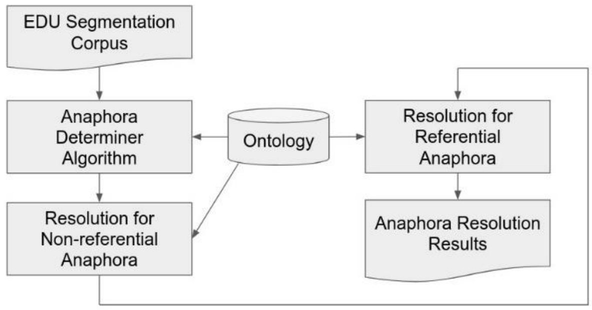

Fig. 1: The overview of the anaphora resolution processes

```
[[สัตว์คล้ายเสือ]<NCM>]<HNpat:Entity:3036>[[ใน]<PRP>]<PRPpat>[[กลุ่ม]<NCA>
[นิมราวีดี]<NCM>]<HNpat:Nom:3037:0>[[ส่วนมาก]<DSO>]<DETpat>[[จะ]<VAX>
[มี]<VRB>]<VRBpat>[[เขี้ยว]<NCM>]<HNpat:Entity:3038>[[บน]<NPP>]<DETpat>
[ที่]<PRL>[@]<Zero:3039:3038>[[มี]<VRB>]<VRBpat>[[ขนาด]<NCA>]-
<HNpat:Entity:3040>[[ยาว]<VAT>[และ]<CON>[ต้น]<VAT>]<ADJpat>
[จน]<SUB>[@]<Zero:3041:3039>[[มองดู]<VRB>]<VRBpat>
[เหมือนกับว่า]<SUB>[@]<Zero:3042:3041>[[มี]<VRB>]<VRBpat>[[ลักษณะ]<NCA>]-
<HNpat:Entity:3043>[[คล้าย]<VPO>[กับ]<PRP>]<PRPpat>[[ดาบ]<NCM>]-
<HNpat:Entity:3044>[[โค้ง]<VAT>[ขนาด]<NCA>[ใหญ่]<VAT>]<ADJpat>
[ส่วน]<SUB>[[เขี้ยว]<NCM>]<HNpat:Elipsis:3045:0>[ที่]<PRL>[[อยู่]<VRB>]-
<VRBpat>[[ด้าน]<CLS>[ล่าง]<NPP>]<DETpat:Entity:3046>
```

Fig. 2: The example of the anaphora tagging in the corpus

Anaphora Determiner Algorithm

The anaphora determiner algorithm is the algorithm to indicate that each phrase in EDU is the entity or the anaphora and also identify the anaphora type to the anaphora. The rule-based is applied to decide to indicate the entity and identify the anaphora type. The anaphora determiner algorithm is shown in Algorithm 1.

The non-recursive phrases that appear in the algorithm are Head Noun (HN pat), Verbal Noun (VNN pat), Time (TIME pat), Classifier (CLS pat), Determiner (DET pat), Adjective (ADJ pat), Amount (AMT pat), transitive Verb (VRB pat) and intransitive Verb (VRI pat). After the anaphora determiner process, the entities and all anaphora will be tagged with the identification number for reference.

Resolution for Non-Referential Anaphora and Referential Anaphora

In this study, the first step to resolving the anaphora is to identify whether the anaphora is non-referential or is referential anaphora. The ranking model by Denis and Baldridge (2008) is selected to resolve the non-referential and also referential anaphora. The ranking model is shown in Eq. 1:

DOI: 10.3844/jcssp.2022.306.315

$$P\left(\phi_i \mid \pi\right) = \frac{\exp\left(\sum_j w_j f_j\left(\pi,\phi_i\right)\right)}{\sum_k \exp\left(\sum_j w_j f_j\left(\pi,\phi_k\right)\right)} \quad (1)$$

Equation 1, π stands for the anaphora type, ϕ_i for the antecedent candidate, f_j for the feature function, w_j for the weight of the feature function, and k for the iterator of all candidates. This equation computes the probability of references given the anaphora type. All anaphora that appear in the training corpus will be evaluated with all features to compute the probability and made the decision. In the training process, the weight adjustment is defined in Eq. 2:

$$w_j = w_j + \alpha\left[f_j\left(\pi,\phi_i\right) - \sum_k P\left(\phi_k \mid \pi\right) f_j\left(\pi,\phi_k\right)\right] \quad (2)$$

Algorithm 1: The anaphora determiner algorithm

input: Q is an array of EDU
begin
  foreach *E in Q* **do**
    if *E has no subject with (VRBpat, VRIpat, ADJpat)* **then**
      Mark Zero at subject
    end
    foreach *H is (HNpat, VNNpat, AMTpat, DETpat) in E* **do**
      if *There is pronoun in H* **then**
        Mark Pronominal
      else if *H is (HNpat, VNNpat) and connect with DETpat* **then**
        Mark Nominal
      else if *H is HNpat and has part-of relation and is a subject* **then**
        if *H follows by preposition of the owner* **then**
          Mark Entity
        else
          Mark Ellipsis
        end
      else if *H is (DETpat, AMTpat) with no (HNpat, TIMEpat, CLSpat, DETpat, ADJpat, VNNpat) before* **then**
        Mark Entity
      else if *H is (HNpat, VNNpat)* **then**
        Mark Entity
      else
        continue
      end
    end
  end
end

Results

Our corpus for training contains a total of 18,248 words and 2,327 EDUs. There are 3,934 entities, 1,272 zero anaphora, 126 nominal anaphora, 64 pronominal anaphora, and 88 ellipses of the owner in the corpus. The precision, recall, and F1 score are used to evaluate the algorithm. The measures are defined as Eq. 3.

Feature Extraction in Non-Referential Anaphora

The features are extracted from the tagged corpus and then store in the database for training purposes. The structure of the feature consists of 3 parts that are feature type, feature value, and weight. Table 1 shows the example of the features of non-referential anaphora in the database.

The feature type and the feature value are encapsulated to the string with the colon connector. The first part of the string is the feature type and the second part is the feature value. The feature type "zero0N4" encapsulated 3 meanings. "zero" means zero anaphora. "0N" means is not non-referential anaphora. And "4" means the fourth kind of feature value. 16 kinds of feature values are used to indicate the non-referential anaphora. Table 2 shows the kinds of feature values for non-referential anaphora.

Verb, syntactic information, and word that surround the anaphora are used as the features for the training model. Due to the non-referential anaphora having no reference, then the only surface word and some syntactic information are considered to be used to indicate the non-referential anaphora.

$$Precision = \frac{\#\,o\;f\;correct\;anaphora\;by\;algorithm}{\#\,o\;f\;anaphora\;determined\;by\;algorithm}$$

$$Recall = \frac{\#\,o\;f\;correct\;anaphora\;by\;algorithm}{\#\,o\;f\;anaphora\;in\,cropus}$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \quad (3)$$

Feature Extraction in Referential Anaphora

The features for referential anaphora are also extracted from the tagged corpus and then stored in the database. The feature structure consists of 4 parts that are feature type, feature value, distance, and weight. Table 3 shows the example of the features of referential anaphora in the database.

Table 1: The example of the features of non-referential anaphora in the database

Features	Weight
elip0N6: CON	1.25
zero0N4: บาง_ที	2.65
pro0N1: ไม่_ได้_ต่อต้าน	1.05
zero0N7: เรียนรู้	1.02
zero0N7: ยัง_ชอบ_กิน	2.61
pro0Y5: CON	1.87

The feature value can be one value or pair value of anaphora and reference. Then, the feature value for referential anaphora can be divided into 3 groups: An anaphora value, a reference value, and pair of anaphora and reference value. The first group is the value of the anaphora and the surrounding information. 16 kinds of the first feature values on the anaphora side are used to indicate the referential anaphora. Table 4 shows the first group of feature values on the anaphora side.

The second group is the value of the reference and the surrounding information. 17 kinds of the second feature values on the reference side are shown in Table 5.

Table 2: The kinds of feature values for non-referential anaphora

1. Verb	2. Verb pos	3. Verb phrase type
4. Word in front	5. Word pos in front	6. Word phrase type in front
7. Word behind	8. Word pos behind	9. Word phrase type behind
10. Syntactic position	11. Head or part of noun	12. Word
13. Pos	14. Phrase type	15. Start paragraph
16. End paragraph		

Table 3: The example of the features of referential anaphora in the database

Features	Weight
zeroXA7: มักจะ_ทำ:1	1.70
elipXB15: ขน:นกกระจอกเทศ	1.09
zeroXC4: ย่อมที่จะ:สำหรับ:1	1.00
proXB10:Dobject:2	1.09
nomXB1: ซ่อน:1	1.00
zeroXB12: นก:8	4.11

Table 4: The first group of feature values on the anaphora side

1. Verb (anaphora): Distance
2. Verb pos (anaphora): Distance
3. Verb phrase type (anaphora): Distance
4. Word in front (anaphora): Distance
5. Word pos in front (anaphora): Distance
6. Word phrase type in front (anaphora): Distance
7. Word behind (anaphora): Distance
8. Word pos behind (anaphora): Distance
9. Word phrase type behind (anaphora): Distance
10. Syntactic position (anaphora): Distance
11. Head or part of a noun (anaphora): Distance
12. Word (anaphora): Distance
13. Pos (anaphora): Distance
14. Phrase type (anaphora): Distance
15. Start paragraph (anaphora): Distance
16. End paragraph (anaphora): Distance

Table 5: The second group of feature values on the reference side

1. Verb (reference): Distance
2. Verb pos (reference): Distance
3. Verb phrase type (reference): Distance
4. Word in front (reference): Distance
5. Word pos in front (reference): Distance
6. Word phrase type in front (reference): Distance
7. Word behind (reference): Distance
8. Word pos behind (reference): Distance
9. Word phrase type behind (reference): Distance
10. Syntactic position (reference): Distance
11. Head or part of a noun (reference): Distance
12. Word (reference): Distance
13. Pos (reference): Distance
14. Phrase type (reference): Distance
15. Word (anaphora): Word (reference)
16. Is-head-word-match: Distance
17. Is-hyponymy: Distance

Table 6: The third group of feature values on both sides of anaphora and reference

1. Verb (anaphora): Verb (reference): Distance
2. Verb pos (anaphora): Verb pos (reference): Distance
3. Verb phrase type (anaphora): Verb phrase type (reference): Distance
4. Word in front (anaphora): Word in front (reference): Distance
5. Word pos in front (anaphora): Word pos in front (reference): Distance
6. Word phrase type in front (anaphora): Word phrase type in front (reference): Distance
7. Word behind (anaphora): Word behind (reference): Distance
8. Word pos behind (anaphora): Word pos behind (reference): Distance
9. Word phrase type behind (anaphora): Word phrase type behind (reference): Distance
10. Syntactic position (anaphora): Syntactic position (reference): Distance
11. Head or part of a noun (anaphora): Head or part of a noun (reference): Distance
12. Word (anaphora): Word (reference): Distance
13. Pos (anaphora): Pos (reference): Distance
14. Phrase type (anaphora): Phrase type (reference): Distance

Table 7: The results of the anaphora resolution

Anaphora Types	Precision	Recall	F1
Zero anaphora (non-referential)	0.66	0.91	0.77
Zero anaphora (referential)	0.78	0.80	0.79
Zero anaphora (overall)	0.75	0.82	0.78
Pronominal anaphora (non-referential)	1.00	1.00	1.00
Pronominal anaphora (referential)	1.00	1.00	1.00
Pronominal anaphora (overall)	1.00	1.00	1.00
Nominal anaphora (non-referential)	1.00	1.00	1.00
Nominal anaphora (referential)	0.96	0.96	0.96
Nominal anaphora (overall)	0.99	0.99	0.99
Ellipsis of the owner (non-referential)	0.70	1.00	0.82
Ellipsis of the owner (referential)	0.87	0.87	0.87
Ellipsis of the owner (overall)	0.84	0.89	0.86
Overall	0.77	0.84	0.81

The third group is the pair value of the anaphora and reference and the surrounding information. 14 kinds of the third feature values on both sides of anaphora and reference are shown in Table 6.

A total of 47 kinds of feature values are used in the resolution for referential anaphora. The distance is set to the maximum of 10 EDUs between the anaphora and the reference. The ranking model is used to find the best probabilistic on the antecedent candidates that are up to 10 EDUs.

Anaphora Resolution Results

The results were evaluated from anaphora determiner, resolution for non-referential anaphora, and resolution for referential anaphora. Each kind of anaphora is evaluated separately and also overall. The results of the anaphora resolution are shown in Table 7.

Zero anaphora is the kind of anaphora that mostly appears in the EDUs. The results show a good precision of 0.75 and a recall of 0.82. The pronominal anaphora is finished with the amazing results that precision is 1.00 and recall is 1.00. These results are successful without using additional knowledge such as gender and number. Because the use of pronominal anaphora in the corpus is not a complicated scenario. Then the only use of the surface word and syntactic information can produce good results in our corpus. The nominal anaphora also recorded high precision of 0.99 and a recall of 0.99. The ontology that provides hyponymy knowledge is useful to resolve the nominal anaphora. The surrounding words in nominal anaphora and reference are also significant to resolving the ranking for nominal anaphora resolution. The ellipsis of the owner recorded high precision of 0.84 and a recall of 0.89. The ontology that provides the meronymy is a significant background knowledge that can be used to identify the entity that is a part of something, especially in the agriculture corpus. The overall results show that the precision is 0.77, the recall is 0.84 and the F1 is 0.81.

Conclusion

In this study, we present the methodology to resolve the anaphora in Thai EDU segmentation. The methodology is done by using the background knowledge to resolve the hyponymy and meronymy relation between the anaphora and the references. The algorithm contains three parts: Anaphora determiner, resolution for non-referential anaphora, and resolution for referential anaphora. The first step is the algorithm to determine the kind of anaphora in each EDU. The algorithm searches each entity in EDU and analyzes the word and the surrounding words together with the ontology to decide the kind of the anaphora. The second step is the resolution for non-referential anaphora. The resolution utilized the ranking model to identify whether anaphora is a non-referential or is a referential anaphora. This resolution works on the only use of the surface word and the surrounding words for learning the model. The final step is the resolution for referential anaphora. The candidate references are generated from the entities in each EDU up to 10 prior EDUs. The ranking model computes the probabilistic value in each candidate and then chooses the candidate with the highest probabilistic value for the referential anaphora. The overall results are that the precision is 0.77, the recall is 0.84 and the F1 score is 0.81. In addition, this study mentions the anaphora types that could be of concern in Thai anaphora resolution especially the ellipsis of the owner. The non-referential anaphora is also significant and could not be overlooked. However, this study is based on the collected corpus that could not be comprehensive. Changing domain can affect the results and might need additional features and also further background knowledge. To ensure the reliability of the results, the making of the comprehensive corpus on various domains and also the modification features could be the focus of future research in this area.

Author's Contributions

Authapon Kongwan: Contributions to the data acquisition, implementation, evaluation, and drafting of the article.

Siti Sakira Kamaruddin: Contributions to planning, designing, reviewing, and approving the article.

Farzana Kabir Ahmad: Contributions to consulting, reviewing the implementation, and approving the article.

Ethics

This article is original and contains unpublished material. The authors have read and approved the manuscript and no ethical issues are involved.

References

Aone, C., & William, S. (1995, June). Evaluating automated and manual acquisition of anaphora resolution strategies. In 33rd Annual Meeting of the Association for Computational Linguistics (pp. 122-129). https://aclanthology.org/P95-1017.pdf

Aroonmanakun, W. (2000). Zero pronoun resolution in Thai: A centering approach. In Burnham, Denis, et. al. Interdisciplinary Approaches to Language Processing: The International Conference on Human and Machine Processing on Human and Machine Processing of Language and Speech. NECTEC: Bangkok (pp. 127-147). http://pioneer.chula.ac.th/~awirote/ling/zero-pronoun-resolution.pdf

Aroonmanakun, W. (2007, December). Thoughts on word and sentence segmentation in Thai. In Proceedings of the Seventh Symposium on Natural language Processing, Pattaya, Thailand, December 13–15 (pp. 85-90).

Boonkwan, P., & Supnithi, T. (2017, June). Bidirectional deep learning of context representation for joint word segmentation and POS tagging. In International Conference on Computer Science, Applied Mathematics and Applications (pp. 184-196). Springer, Cham. https://link.springer.com/chapter/10.1007/978-3-319-61911-8_17

Carlson, L., Marcu, D., & Okurowski, M. E. (2003). Building a discourse-tagged corpus in the framework of rhetorical structure theory. In Current and new directions in discourse and dialogue (pp. 85-112). Springer, Dordrecht. https://link.springer.com/chapter/10.1007/978-94-010-0019-2_5

Clark, K., & Manning, C. D. (2016a). Deep reinforcement learning for mention-ranking coreference models. arXiv preprint arXiv:1609.08667. https://arxiv.org/abs/1609.08667

Clark, K., & Manning, C. D. (2016b). Improving coreference resolution by learning entity-level distributed representations. arXiv preprint arXiv:1606.01323. https://arxiv.org/abs/1606.01323

Cohen, W. W., & Singer, Y. (1999). A simple, fast and effective rule learner. AAAI/IAAI, 99(335-342), 3. https://www.aaai.org/Papers/AAAI/1999/AAAI99-049.pdf?ref=https://githubhelp.com

Daelemans, W., Zavrel, J., Van Der Sloot, K., & Van den Bosch, A. (2004). Timbl: Tilburg memory-based learner. Tilburg University. file:///C:/Users/User/Downloads/Timbl_6.4_Manual.pdf

Denis, P., & Baldridge, J. (2008, October). Specialized models and ranking for coreference resolution. In Proceedings of the 2008 conference on empirical methods in natural language processing (pp. 660-669). https://aclanthology.org/D08-1069.pdf

Di Eugenio, B. (1998). Centering in italian. Centering theory in discourse, 115-137. ISBN-10: 9780198236870.

Durrett, G., & Klein, D. (2013, October). Easy victories and uphill battles in coreference resolution. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (pp. 1971-1982). https://aclanthology.org/D13-1203.pdf

Ge, N., Hale, J., & Charniak, E. (1998). A statistical approach to anaphora resolution. In Sixth workshop on very large corpora. https://aclanthology.org/W98-1119.pdf

Grosz, B. J., Joshi, A. K., & Weinstein, S. (1995). Centering: A framework for modelling the local coherence of discourse. https://repository.upenn.edu/ircs_reports/116/

Hobbs, J. R. (1978). Resolving pronoun references. Lingua, 44(4), 311-338. doi.org/10.1016/0024-3841(78)90006-2

Iida, M. (1996). Discourse coherence and shifting centers in Japanese texts. arXiv preprint cmp-lg/9609007. https://arxiv.org/abs/cmp-lg/9609007

Ketui, N., Theeramunkong, T., & Onsuwan, C. (2013). Thai elementary discourse unit analysis and syntactic-based segmentation. International Information Institute (Tokyo). Information, 16(10), 7423. http://203.131.210.88/larts/file/Jpdf_1.pdf

Ketui, N., Theeramunkong, T., & Onsuwan, C. (2015). An edu-based approach for thai multi-document summarization and its application. ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), 14(1), 1-26. https://dl.acm.org/doi/abs/10.1145/2641567

Kongwan, A., Kamaruddin, S. S. B., & Ahmad, F. B. K. (2020). Thai EDU Segmentation Using Clue Markers and Syntactic Information from Shallow Parser. Journal of Theoretical and Applied Information Technology, 98(18), 3853-3869. http://www.jatit.org/volumes/Vol98No18/12Vol98No18.pdf

Kongyoung, S., Rugchatjaroen, A., & Kosawat, K. (2015, November). TLex+: a hybrid method using conditional random fields and dictionaries for Thai word segmentation. In International Conference on Knowledge, Information and Creativity Support Systems (pp. 112-125). Springer, Cham. https://link.springer.com/chapter/10.1007/978-3-319-70019-9_10

Lappin, S., & Leass, H. J. (1994). An algorithm for pronominal anaphora resolution. Computational linguistics, 20(4), 535-561. https://aclanthology.org/J94-4002.pdf

Lee, H., Chang, A., Peirsman, Y., Chambers, N., Surdeanu, M., & Jurafsky, D. (2013). Deterministic coreference resolution based on entity-centric, precision-ranked rules. Computational linguistics, 39(4), 885-916. https://direct.mit.edu/coli/article/39/4/885/1463/Deterministic-Coreference-Resolution-Based-on

Lee, H., Surdeanu, M., & Jurafsky, D. (2017a). A scaffolding approach to coreference resolution integrating statistical and rule-based models. Natural Language Engineering, 23(5), 733-762. doi.org/10.1017/S1351324917000109

Lee, K., He, L., Lewis, M., & Zettlemoyer, L. (2017b). End-to-end neural coreference resolution. arXiv preprint arXiv:1707.07045. https://arxiv.org/abs/1707.07045

Liu, L., Song, Z., & Zheng, X. (2020). Improving coreference resolution by leveraging entity-centric features with graph neural networks and second-order inference. arXiv preprint arXiv:2009.04639. https://arxiv.org/abs/2009.04639

Luo, X. (2005, October). On coreference resolution performance metrics. In Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (pp. 25-32). https://aclanthology.org/H05-1004.pdf

Marcu, D. (1998). A surface-based approach to identifying discourse markers and elementary textual units in unrestricted texts. In Discourse Relations and Discourse Markers. https://aclanthology.org/W98-0301.pdf

Marcu, D. (1999, June). A decision-based approach to rhetorical parsing. In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (pp. 365-372). https://aclanthology.org/P99-1047.pdf

McCallum, A., & Wellner, B. (2004). Conditional models of identity uncertainty with application to noun coreference. Advances in neural information processing systems, 17. https://proceedings.neurips.cc/paper/2004/hash/1680829293f2a8541efa2647a0290f88-Abstract.html

Mitkov, R., Evans, R., & Orasan, C. (2002, February). A new, fully automatic version of Mitkov's knowledge-poor pronoun resolution method. In International Conference on Intelligent Text Processing and Computational Linguistics (pp. 168-186). Springer, Berlin, Heidelberg. https://link.springer.com/chapter/10.1007/3-540-45715-1_15

Netisopakul, P., & Wohlgenannt, G. (2017, July). The state of knowledge extraction from text for thai language. In 2017 6th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI) (pp. 379-384). IEEE. https://ieeexplore.ieee.org/abstract/document/8113274/

Netisopakul, P., & Wohlgenannt, G. (2018). A survey of Thai knowledge extraction for the semantic web research and tools. IEICE TRANSACTIONS on Information and Systems, 101(4), 986-1002. https://search.ieice.org/bin/summary.php?id=e101-d_4_986

Ng, V., & Cardie, C. (2002, July). Improving machine learning approaches to coreference resolution. In Proceedings of the 40th annual meeting of the Association for Computational Linguistics (pp. 104-111). https://aclanthology.org/P02-1014.pdf

DOI: 10.3844/jcssp.2022.306.315

Nicolae, C., & Nicolae, G. (2006, July). Best cut: A graph algorithm for coreference resolution. In Proceedings of the 2006 conference on empirical methods in natural language processing (pp. 275-283). https://aclanthology.org/W06-1633.pdf

Pathanasin, S. (2018). Coherence in Thai Students' Essays: An Analysis using Centering Theory. Manusya: Journal of Humanities, 21(2), 112-130. https://brill.com/view/journals/mnya/21/2/article-p112_6.xml

Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543). https://aclanthology.org/D14-1162.pdf

Poesio, M., Stuckardt, R., Versley, Y., & Vieira, R. (2016). Early approaches to anaphora resolution: Theoretically inspired and heuristic-based. In Anaphora Resolution (pp. 55-94). Springer, Berlin, Heidelberg. https://link.springer.com/chapter/10.1007/978-3-662-47909-4_3

Slayden, G., Hwang, M. Y., & Schwartz, L. (2010, August). Thai sentence-breaking for large-scale SMT. In Proceedings of the 1st Workshop on South and Southeast Asian Natural Language Processing (pp. 8-16). https://aclanthology.org/W10-3602.pdf

Soon, W. M., Ng, H. T., & Lim, D. C. Y. (2001). A machine learning approach to coreference resolution of noun phrases. Computational linguistics, 27(4), 521-544. https://direct.mit.edu/coli/article/27/4/521/1748/A-Machine-Learning-Approach-to-Coreference

Strube, M., & Hahn, U. (1996). Functional centering. In 34th Annual Meeting of the Association for Computational Linguistics, pages 270–277. https://dl.acm.org/doi/abs/10.3115/981863.981899

Sukthanker, R., Poria, S., Cambria, E., & Thirunavukarasu, R. (2020). Anaphora and coreference resolution: A review. Information Fusion, 59, 139-162. doi.org/10.1016/j.inffus.2020.01.010

Sukvaree, T., Kawtrakul, A., & Caelen, J. (2007, August). Thai text coherence structuring with coordinating and subordinating relations for text summarization. In International and Interdisciplinary Conference on Modeling and Using Context (pp. 453-466). Springer, Berlin, Heidelberg. https://link.springer.com/chapter/10.1007/978-3-540-74255-5_34

Zeldes, A., & Zhang, S. (2016, June). When annotation schemes change rules help: A configurable approach to coreference resolution beyond OntoNotes. In Proceedings of the Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2016) (pp. 92-101). https://aclanthology.org/W16-0713.pdf

Zhou, N., Aw, A., Lertcheva, N., & Wang, X. (2016, December). A word labeling approach to Thai sentence boundary detection and POS tagging. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers (pp. 319-327). https://aclanthology.org/C16-1031/