Product Name Identification and Classification in Thai Economic News

Nattadaporn Lertcheva
Department of Linguistics
Chulalongkorn University
nattadaporn@gmail.com

Wirote Aroonmanakun
Department of Linguistics
Chulalongkorn University
awirote@chula.ac.th

Abstract

The purpose of this research is to analyze the patterns of the product names used in Thai economic news and to find clues that could be used to identify the product names' boundaries and their categories. It is found that the patterns of Thai product names are quite varied. Thirty two patterns are found in this study. While some clues like collocation and the context of names can be used for identifying product names, many of them cannot be identified by these means. This indicates that the task of product named entity recognition is an interesting task for Thai language processing.

1 Introduction

Most named entity recognition research has been focused on person, location, and organization names. Though other proper names, such as biomedical names and product names, are important in language processing, only a little research has been done on recognizing these names in Thai such as Lertcheva and Aroonmanakun (2009). Since different types of names have different patterns and characteristics, basic linguistic knowledge of the names is needed for imposing any rules or features for any rule-based or statistical-based named entity recognition systems. This paper presents basic knowledge of Thai product names. A corpus of Thai economic news is used in analyzing product names. Patterns and variations of their forms in texts are analyzed. In this paper, background information of product names and relevant research will be presented first. Then, the corpus and annotation used in marking Thai product names will be described in section 3. The results of the analysis will be presented in sections 4 and 5 followed by the conclusion.

2 Background Knowledge

Unlike a person name, an organization name, or a location name, which is normally used to refer to one unique referent, a product name is used to refer to many referents categorized under the same product. Product names are a kind of proper name because each is created to refer to a certain product produced by a company. This section describes the definition of product names and product categories used in this study. Although product named entity recognition has been analyzed in Lertcheva and Aroonmanakun (2009) which focused on linguistic analysis of the product names for solving product name identification, this paper furthers the study by analyzing product names in detail using a larger corpus. Moreover, we will propose the pattern of product names and describe the components used to classify different types of product.

2.1 Definition of Product Names

To distinguish one product from the same products produced by other companies, trademarks or brand names are usually used in the product names. However, previous research used the terms "product names"with different meanings. For example, Liu et al. (2005) defined a product name as a name consisting of a trade mark and product type, e.g. Nokia 3310. Nilsson and Malmgren (2005) defined a product name as a term under brand names. In other words, a brand name consists of a trademark, a product name, and a service name. Trademarks have a broader scope than product names or service names. For example, Volvo is regarded as a trademark while Volvo C70 is considered a product name. Boonpaisarnsatit (2005) used the

Proceedings of the 2011 Named Entities Workshop, IJCNLP 2011, pages 58–64,
Chiang Mai, Thailand, November 12, 2011.

term “product names” differently from Liu et al. (2005) and Nilsson and Malmgren (2005). What is called “product name” in Boonpaisarnsatit (2005) is actually a generic noun indicating a category of product. He referred to “brand names” as the combination of product name and trademark. For example, รถยนต์|โตโยต้า is analyzed as consisting of a product name รถยนต์-‘car’ and a trademark โตโยต้า-‘Toyota’. The use of a generic noun when referring to a product is a characteristic of referring to products in Thai. In this study, we use the term product name as defined in Lertcheva and Aroonmanakun (2009) which is a linguistic expression consisting of a generic noun, a brand name indicator, a brand name, a product type indicator, and a product type.

2.2 Product Categories

In product named entity recognition, the task includes not only identifying the boundary but also the type of the product. However, there is no standard classification of product category. In this paper, we use the classification listed by the Department of Export Promotion, Ministry of Commerce of Thailand and Wikipedia as a basis of classification and divide the products into 26 categories as follows:

1. Foods
2. Medical devices
3. Pharmaceutical
4. Cosmetic and spa products
5. Eyewear brands
6. Electrical products and parts / Electronics
7. Automotive / auto parts and accessories
8. Building materials and hardware items
9. Chemicals and plastic resins
10. Printing products, paper and packaging
11. Machinery and equipment
12. Gems and jewelry
13. Watches/Clocks
14. Bags/Footwear/Leather Products
15. Textiles, garments and fashion accessories
16. Sporting goods
17. Furniture and parts
18. Gift and decorative items/handicrafts
19. Household products
20. Home textiles
21. Toys and games
22. Stationery/Office supplies and Equipment
23. Tobacco
24. Farming products
25. Cleaning products
26. Miscellaneous

3 Corpus and Annotation

To reveal patterns of product names in Thai, a corpus of Thai economic news is used. The corpus size is 178,474 words, in which 2,463 product names are found.[1] Since the language used in the headlines usually has different style from the body text, in this study, we analyze only the product names found in the body text of the news. TEI annotation style is used in marking up product names. A product name is tagged by using<productNametype=“Product’s_Category” >…</productName>. The annotation of the components in product names is as follows.

1. <genericNoun>.....</genericNoun> is used for tagging words used to describe the type of product. For example, **โทรศัพท์มือถือ**|โนเกีย consists of a compound noun, โทรศัพท์มือถือ-‘mobile phone’, and a brand name “Nokia”. Although the corpus is collected from Thai economic news, generic nouns are not always written in Thai script. Even though English names can be transliterated using Thai script, they are often written in English. For example, the product name “LCD TV รุ่นAN-LT 322 DU" begins with a generic noun in English “LCD TV” followed by a product type indicator in Thai รุ่น-‘model’ and then the product type in English “AN-LT 322 DU”. Generic nouns can be a simple word, a compound, or a phrase e.g.อาหารทะเลแช่แข็ง-‘frozen sea food’.

2. <brandIndicator>.....</brandIndicator> is used to mark a brand indicator, or a word indicating the brand name. Brand indicators found in the corpus are limited to words like ตรา-‘brand’, ยี่ห้อ-‘brand’, ตระกูล-‘family’, เครื่องหมายการค้า-‘trademark’, ชื่อ-‘name’, ผลิตภัณฑ์-‘product’, and แบรนด์-‘brand’. Brand indicators can be preceded by some prepositions like ภายใต้-‘under’, e.g. ภายใต้|ผลิตภัณฑ์ = ‘under’+‘product’, or it can be modified by an adjective like ใหม่-‘new’, e.g. แบรนด์|ใหม่= ‘brand’+‘new’.

3. <brandName>.....</brandName> is used to mark the brand name of the product. The brand name is normally a trademark named for the products. Brand names are sometimes found

[1] The corpus can be downloaded from http://pioneer.chula.ac.th/~awirote/Data-Nattadaporn.zip

written in English, such as, เครื่องสำอาง|**DHC** = a generic noun 'cosmetic' + a brand name '**DHC**'

4. <proIndicator>..... </proIndicator> is the markup for the product type indicator used to identify the product type. Product type indicators found in the corpus are รุ่น-'type', ซีรีย์-'series', สูตร-'formula', กลิ่น-'scent', รส/รสชาติ-'taste', ชนิด-'type',ครอบครัวตระกูล-'family'. These product type indicators sometimes can be modified by an adjective, such as รุ่น+ใหม่= 'type'+'**new**'.

5. <productType>.....</productType> is for tagging product subtype under the same brand name. It is found that either common nouns or proper nouns can be used as a product type. In food product names, a common noun related to taste is likely to be used indicating its subtype, e.g. มาม่า+รส+**ต้มยำกุ้ง**– 'Mama'+'taste'+'**spicy lemongrass with shrimp**'. For technology products, a proper noun is usually used to identify the subtype, e.g. the name ยาริส-'Yaris' is used to indicate a specific model of the car, โตโยต้า+ยาริส–'Toyota'+ 'Yaris'

4 Product Name Identification

Product names in Thai consist of five components as stated in the previous section. However, the patterns can be varied. To identify a product name, its patterns and contextual clues have to be examined. In this study, we found 32 patterns of product names. These patterns can be categorized into 4 groups, head only, head-initial, head-centre, and head-final (section 4.1). Then, a study of context clues for identifying product names is presented in section 4.2.

4.1 Pattern of Product Names

Of the 32 patterns, brand name and product type are the core part of the product name. A brand name is used to distinguish the product from the same one produced by other companies. A product type is usually used to differentiate similar products under the same brand name. Every pattern of product name would have the brand name as its core part. If the brand name is omitted, the product type would be used as the core part of the product name. These two components are essential in uniquely identifying the product. Therefore, 'head' in this paper refers to a brand name or a product type.

The symbols used in the pattern of product names are described as follows.

1. (…) indicates the component that can be omitted in the product name.

Example: A + (B) + C = A + B + C or A + C

2. […] indicates that the component is required in the pattern.

Example: [+brand] means that a brand indicator must be present in this pattern and must be the word 'brand'.

3. {…} indicates that at least one element in the braces must be present.

Example: {A + B} + C = A + C or B + C or A + B + C

4. | is used for marking the selection of only one choice.

Example: A|B + D = A + D or B + D

From the 32 patterns found in the 2,463 product names, we can categorize them into 4 groups as follows:

1. Head Structures

This pattern consists of one component, brand name or product type, functioning as the head word. From all the product names, the pattern with the brand name as head is found in 39.26% of the product names while the pattern with product type as head is found in 4.06% of the product names.

- Brand name

This pattern is found when the product name is used continuously in the text or in an illustration sentence. For example, <product Name type="cosSpa" ID="P03"><brandName> จุยซ์บิวตี้</brandName></productName> is a name consisting of only the brand name "Juice Beauty".

- Product type

This pattern is found when the product name is continuously referred to in the text. The product type can be either a common noun or a proper noun. For example, <product Nametype ="Elec"><productType>ธิงค์แพดเอ็กซ์100อี</product Type></productName>has a proper name as the product type, "ThinkPad X 100E." In the example, <productNametype="food"><product Type>หมูสับ</productType></productName>, the product type is a common noun referring to "minced pork". This pattern, in which only a common word functions as the product type, is acceptable only if the same product is previously referred to using a product name pattern containing a brand name. This is because, unlike a proper noun, a common noun by itself cannot specify what the product is. For example, we can use the product type "Jazz" without mentioning a brand name because the reader can understand what we are referring to. In contrast, we cannot use a common word likeหมูสับ – "minced pork" as

the product name when first introduced in the text since the readers cannot understand what the product is.

2. Head-Initial Structures

This is the pattern in which the head is located at the beginning. This pattern consists of 4 sub-patterns which account for 10.19% of the product names.

▪ **Brand name** + {brand name indicator [+brand] + generic noun }

Example: <productName type="gems"><brandName>ดามิอานิ</brandName><brandIndicator>แบรนด์</brandIndicator><genericNoun>เครื่องประดับ</genericNoun></productName>

This example consists of a brand name “Damiani”, a brand indicator “brand” and a generic noun “jewelry”.

▪ **Brand name** + {generic noun + product type indicator}+ product type

Example: <productName type=“Elec” ID=“P02”> <brandName> แบล็กเบอรี่ </brand Name> <proIndicator>รุ่น</proIndicator><productType> โบลด์</productType> </productName>

This example consists of a brand name “Blackberry”, a product type indicator “type” and a product type “Bold”.

▪ **Brand name** + product type + (generic noun)

Example: <productName type="food"> <brandName>ไวตามิ้ลค์</brandName> <productType>โลว์ชูการ์</productType> </productName>

This example consists of a brand name “Vitamilk” and a product type “Low sugar”.

▪ **Product type** + product type indicator

Example: <productName type=“Elec”> <productType>ธิงค์แพด</productType> <proIndicator>ซีรีส์</proIndicator> </productName>

This example consists of a product type “ThinkPad” and a product type indicator “series”.

3. Head-Centre Structures

This is the pattern in which the head is located at the centre of the structure. This pattern consists of 5 sub-patterns which account for 5.08% of the product names.

▪ Generic noun | brand name indicator + **brand name** + generic noun

Example: <productName type="food" ID=“P02”><brandIndicator>แบรนด์</brandIndicator><brandName>อาร์ทรี</brandName><genericNoun>ชาพร้อมดื่ม</genericNoun></productName>

This example consists of a brand indicator “brand”, a brand name “Artea” and a generic noun “tea”.

▪ {Generic noun + brand name indicator | product type indicator}+ **brand name** + product type

Example: <productName type="cosSpa"> <genericNoun>ยาสีฟัน</genericNoun> <brandIndicator>ยี่ห้อ</brandIndicator> <brandName>ฟลูโอคารีล</brandName> <productType>40 พลัส</productType> </productName>

This example consists of ageneric noun “toothpaste”, a brand indicator “brand”, a brand name “Fluocaril” and a product type “40 plus”.

▪ Generic noun + **brand name** + product type + generic noun

Example:<productName type=“Auto”> <genericNoun>รถ</genericNoun><brandName> เชฟโรเลต</brandName><productType>โคโลราโด </productType><genericNoun>ปิคอัพอเมริกันพันธุ์แกร่ง </genericNoun></productName>

This example consists of a generic noun “car”, a brand name “Chevrolet”, a product type “Colorado” and a generic noun “American pick-up”.

▪ Brand name indicator + **brand name** + brand name indicator + generic noun

Example: <productName type="fashion"> <brandIndicator>ไฟติ้งแบรนด์ชื่อ</brandIndicator> <brandName>จีแอนด์จี</brandName> (Guy&Girl) <brandIndicator>แบรนด์</brandIndicator> <genericNoun>ชุดชั้นใน</genericNoun> </productName>

This example consists of a brand indicator “fighting brand”, a brand name “G&G”, a brand indicator “brand” and a generic noun “underwear”.

▪ Generic noun + (brand name indicator) + **brand name** + (product type) + product type indicator + product type

Example: <productName type=“Auto”> <genericNoun>รถ</genericNoun> <brandName>ฮอนด้า</brandName> <productType>ซิตี้</productType><proIndicator> รุ่น</proIndicator><productType>ปี2008 </productType></productName>

This example consists of a generic noun “car”, a brand name “Honda”, a product type “City” a product type indicator “type” and a product type “year 2008”

4. Head-Final Structures

Besides the pattern head only structure, this is the most commonly used structure in product names. The pattern has the head located at the final part of the structure. This pattern consists of 4 sub-patterns which account for 41.41% of the product names.

▪ (generic noun) + brand name indicator + **brand name**

Example: <productName type=“Elec”> <genericNoun>โทรศัพท์เคลื่อนที่</genericNoun>

<brandIndicator>ภายใต้แบรนด์</brandIndicator>

<brandName>แบล็กเบอร์รี่</brandName>

</productName>

This example consists of a generic noun “mobile phone”, a brand indicator “under brand” and a brand name “Blackberry”.

▪ (generic noun) + brand name indicator + generic noun + brand name indicator + **brand name**

Example: <productName type="food" ID="P01"> <genericNoun>ข้าวสารบรรจุถุง</genericNoun>

<brandIndicator>ภายใต้แบรนด์</brandIndicator>

<genericNoun>ข้าว</genericNoun>

<brandIndicator>ตรา</brandIndicator>

<brandName>ฉัตร</brandName></productName>

This example consists of a generic noun “a bag of rice”, a brand indicator “under brand”, a generic noun “rice”, a brand indicator “brand” and a brand name “Chut”

▪ (brand name indicator [+brand]) + generic noun + **brand name**

Example:<productName type="food"> <brandIndicator>แบรนด์</brandIndicator>

<genericNoun>น้ำผลไม้</genericNoun>

<brandName>แบรี่</brandName></productName>

This example consists of a brand indicator “brand”, a generic noun “juice” and a brand name “Berri”.

▪ {Generic noun + product type indicator} + **product type**

Example: <productName type=“Auto”> <proIndicator>รุ่น</proIndicator> <productType> ซีรีส์ 7 ซีดาน</productType> </productName>

This example consists of a product type indicator “type” and a product type “Series 7 Sedan”.

Thai product names tend to be used with head structure and head-final respectively. Head-structure can be used without causing any confusion because normally the product is previously referred to in the text. The preference for the head-final structure conforms to the structure of a proper name in Thai, in which a proper name is preceded by a common noun indicating its class, e.g. โรงเรียน|สวนกุหลาบ= school+ ‘Suankularp’, วัด|บัวขวัญ= temple+‘Buakhwan’, etc. Therefore, readers will perceive the kind of product before the name of products. e.g., ปลาราด พริก|ตรา|ปลายิ้ม = fish with a chili sauce + a brand indicator ‘brand’ + a brand name ‘PlaYim’

4.2 Clues for Identifying Product Names

To find contextual clues that would be useful in identifying product names, words collocated with the product names and specific sentence patterns are examined as follows:

1. Word collocations

This section emphasizes the study of words collocated with the product names. A preliminary observation shows that some words located in front of product names tend to have a meaning related to products such as ‘seller’, ‘buyer’, ‘importer’, ‘sell’, ‘produce’, ‘import’ etc. To determine the efficacy of these words as an indicator of the product names, we analyzed the occurrence of every word found in front of a product name within the span of four words. Words occurring in the corpus less than 6 times were excluded. Then, a percentage of how often the words collocated with product names was calculated and sorted. In this study, words with more than 50% co-occurrence with a product name are considered useful. Only three words are found with this criterion. They are ผู้ผลิต - ‘a producer’, แนะนำ - ‘introduce’ and ผู้แทนจำหน่าย - ‘a dealer’. When the span is set to be three words before the product name, only two words are found useful, namely แนะนำ - ‘introduce’ and ผู้แทนจำหน่าย - ‘a dealer’.

Although a preliminary observation intuitively indicates the close relation between the product name and its collocations, the result does not confirm that observation because the percentages of co-occurrences for most of the collocates are lower than 50%.

2. Illustration sentences

A sentence pattern that is found to be useful for identifying a product name is the sentence with illustration. In this pattern, product names are found as a list of illustrations after the words ได้แก่ - ‘for example’,and เช่น- ‘such as’. The last

product name usually comes after the conjunction และ- ‘and’. In this example, ผู้จัดหาเสื้อผ้า|แบรนด์ |**เช่น**|ลีวายส์ |และ |แรงเลอร์ (clothing dealer + brand + **such as** + Levi’s + and + Wrangler), two product names are listed after the word เช่น-‘such as’.

5 Product Category Identification

The task of product named entity recognition includes not only identifying product name boundaries but also product categories. In this section, we describe the criteria used for identifying product categories. From 2,463 product names, we found that only 1,603 product names (65%) can be assigned to a product category by considering either the components in the product name or contextual clues.

1. Components in the product name

Of those 1,603 names, the product categories can be determined for 1,172 by considering the components within the product names. Components that are useful are generic nouns, brand names, and product types.

- Generic noun

Product categories can be easily determined from the generic noun in the product name. For example,วิทยุ|โซนี่ = **a radio** + a brand name ‘Sony’ is categorized as ‘Electrical products’ because ‘radio’ is a subclass of electrical products. In this example,น้ำดื่ม|สิงห์ = **drinking water** + a brand name ‘Singha’ is categorized as ‘Foods’ because ‘drinking water’ is a subtype of food.

- Brand name

For some names, a part of the brand name can be useful in identifying its category. For example, the brand name ไวตามิลค์ (Vita**milk**) is categorized as ‘Foods’ because there is a word ‘milk’ within the brand name. In this example, ไอโฟน (i**phone**) is categorized as ‘Electronic products’ because of the word ‘phone.’ The brand name เนสท์กาแฟ (Nescafé) is used to categorize the product as ‘Foods’ because the word ‘café’ in Thai means coffee.

- Product type

In some cases, product category can be inferred from the product type. For example, มาม่า|รส|หมูสับ= a brand name ‘Mama’ + a product type indicator ‘taste’ + a product type **‘minced pork’** can be categorized as ‘Foods’ because of the product type ‘minced pork’.

2. Contextual clues

When components in the product name cannot be used to identify the product category, a contextual clue, which comes from a previous mention of the product name in the text, is used. It is found that the categories for 431 product names can be identified by referring back to the same product names previously presented in the text. If a product is referred to more than once in the text, its category is usually identified by considering the components inside the first mention of the name. When the same product is referred to again using a reduced form, its category can be inferred from the previous mention.

In sum, based on the analysis of 2,463 product names, we found that categories can be identified for only 65% of them by analyzing the components inside the product name (1,172) or by referring to a previous mention of the product name (431). The rest, 860 product names (35%), cannot be assigned to their categories using these means. It seems that background knowledge is needed in identifying the product category. These are usually a product which is well known, e.g. โค้ก= ‘Coke’, แพนทีน= ‘Pantene’, etc. Thus, product category identification is not an easy task.

6 Conclusion

This study concerns both product name and product category identification. A linguistic analysis of Thai product names is carried out to reveal patterns of product names and clues that would be useful for product named entity recognition in Thai.

Though there is some preference for the head-only and head-final structures in Thai product names, it is found that the patterns of Thai product names are quite varied. In addition, there is no explicit clue for identifying a product name. Using collocates alone seems to be insufficient for identifying the product name.

For product category identification, some inner clues can be found from the components in the product names. Keeping track of products referred to in the discourse can also help in identifying the category when the name is used in a reduced form. However, categories cannot be identified for a number of product names by this means.

Therefore, the problem of Thai product named entity recognition is not an easy task. Further research on this topic is needed. A general named entity recognition model should be

implemented to verify whether the model that has been used in Thai named entity recognition could resolve this problem. We think that a named entity recognition that uses both word forms and part-of-speech sequences should suffice for identifying the product name boundaries. But identifying product category, if it is needed, should be implemented separately by keeping track of product names found previously and creating semantic relations between the product names and contextual words.

Acknowledgments
This research is supported by The Thailand Research Fund (TRF) under grant no MSG53Z0007, and partially supported by Chulalongkorn University Centenary Academic Development Project.

References

Boonpaisarnsatit, N. 2005.*Semantic analysis of Thai Products' Brand names*.Unpublished master's thesis, Chiang Mai University, Thailand.

Department of Export Promotion.Ministry of Commerce.*Product's information*.Retrieved from: http://www.depthai.go.th[accessed 11 March 2009]

Lertcheva, N. and Aroonmanakun, W. 2009.A Linguistic Study of Product Names in Thai Economic News.In *Proceeding of the 8th international symposium on natural language processing*. October 20-21, 2009. Bangkok. Thailand

Liu, F., Zhao, J., Lv, B., Xu, B., and Yu, H. 2005.Product Named entity Recognition Based on Hierarchical Hidden Markov Model.In *Proceedings of the 4thSIGHAN Workshop on Chinese Language Processing*.

Nilsson, K., and Malmgren, A. 2005. Towards automatic recognition of product names: An exploratory study of brand names in economic texts. In *Proceedings of the15th NODALIDA conference*, Joensuu.

Settels, B. 2004.*Biomedical Named Entity Recognition Using Conditional Random Fields and Rich Feature Sets*.

Wikipedia.*Category:Brands by product type*. Retrieved from: http://en.wikipedia.org/wiki/Category:Brands_by_product_type [accessed 25 December 2008]