A Corpus Study and Annotation Schema for Named Entity Recognition and Relation Extraction of Business Products

Saskia Schön, Veselina Mironova, Aleksandra Gabryszak, Leonhard Hennig
DFKI GmbH
Berlin, Germany
{firstname.lastname}@dfki.de

Abstract
Recognizing non-standard entity types and relations, such as B2B products, product classes and their producers, in news and forum texts is important in application areas such as supply chain monitoring and market research. However, there is a decided lack of annotated corpora and annotation guidelines in this domain. In this work, we present a corpus study, an annotation schema and associated guidelines, for the annotation of product entity and company-product relation mentions. We find that although product mentions are often realized as noun phrases, defining their exact extent is difficult due to high boundary ambiguity and the broad syntactic and semantic variety of their surface realizations. We also describe our ongoing annotation effort, and present a preliminary corpus of English web and social media documents annotated according to the proposed guidelines.

Keywords: Named Entity Recognition, Relation Extraction, Information Extraction

1. Introduction
Recognizing non-standard entity and relation types is an important task in many real-world information extraction applications like relation extraction, knowledge base construction and question answering. In areas such as market research and supply chain management, many companies would benefit from systems that automatically and continuously acquire up-to-date information about producers, vendors and other suppliers of specific parts, products, new technologies and components. Similarly, the construction of knowledge graphs that store supplier and vendor relationships would clearly benefit from information extraction approaches by reducing the manual effort required to create and maintain such databases. For example, in both scenarios it would be useful to extract information about e.g. a CompanyProvidesProduct relation from a news text like “Sensata Technologies’ products include speed sensors, motor protectors, and magnetic-hydraulic circuit breakers”, where the product argument refers to a non-consumer product or product class entity such as “speed sensors” or “magnetic-hydraulic circuit breakers”.

However, when it comes to such specific domains, developing named entity recognition algorithms is severely hampered by the lack of publicly available training data and the difficulty of accessing existing dictionary-type resources, such as product catalogs. Many available named entity recognition corpora consist of general news articles (Tjong Kim Sang and De Meulder, 2003; Doddington et al., 2004; Weischedel et al., 2013), while information about B2B products is typically available on non-journalistic, specialized web portals and forums. Product mentions, as in the example above, are often general noun phrases, instead of proper names, which increases the difficulty of detecting them using gazetteer-based approaches. In addition, relational information about companies and their products is very limited in freely available knowledge bases (KB), such as Freebase (Bollacker et al., 2008), Wikidata (Vrandečić and Krötzsch, 2014), or DBpedia (Auer et al., 2007), since these KBs are in large parts based on Wikipedia, which aims to exclude commercial, non-encyclopedic information. For example, DBpedia contains only approximately 60,000 triples for the CompanyProvidesProduct relation.

To address these problems, and to gain a better understanding of product mentions and their linguistic properties, in this study we first collect a large number of noisy product mentions. This is achieved with a bootstrapping approach that uses a set of manually defined lexical patterns for the relation CompanyProvidesProduct (Section 3.). We analyze the resulting set of mentions, and find that they often include extraneous linguistic material that should not be considered a part of the product extent, such as prepositional phrases and appositions. Consequently, we develop an annotation schema for product mentions and the CompanyProvidesProduct relation, in order to guide the manual annotation of texts (Section 4.). We are currently building a corpus of English web and social media documents with annotations for product entity and CompanyProvidesProduct relation mentions based on these annotation guidelines. We will make a first version of the corpus available to the community (Section 6.). The overall goal of our work is to make extraction of non-standard, B2B products and relations from unstructured text easier and more reliable.

2. Related Work
Most research in Named Entity Recognition (NER) has focused on common entity types, such as persons, organizations, and locations (Tjong Kim Sang and De Meulder, 2003; Finkel et al., 2005; Derczynski et al., 2016), and numeric types like date and time expressions (Ströten and Gertz, 2013). Only a few corpora cover other entity types, such as geopolitical entities and facilities (Doddington et al., 2004; Weischedel et al., 2013). Corpora that include product annotations are rare: the BBN corpus covers (consumer) products mentioned in Wall Street Journal news articles (Weischedel and Brunstein, 2005). Liu et al. (2011) describe a corpus of tweets that has been annotated with products, but the dataset is not publicly avail-
able. Recent research in fine-grained NER has produced distantly (Ling and Weld, 2012; Weischedel et al., 2013) or weakly supervised datasets (Ni et al., 2017) using Freebase and Wikipedia, which therefore inherit the coverage and specificity limitations of these resources. The datasets include products and their subtypes, but the entities are generally consumer products, such as cars, mobile phones, and games. In the case of Ni et al. (2017), the dataset is not publicly available.

For relation extraction, there exist only very few datasets that have been manually annotated with company-product-related information. FrameNet (Baker et al., 1998) contains example sentences marked up with frames that provide information that is similar to the companyProvidesProduct relation, such as the Business, Commerce_sell and Manufacturing frames. The SemEval 2010 Task 8 contains 968 sentences annotated with pairs of nominals for mentions of the Product-Producer relation (Hendrickx et al., 2010). While the relation’s name suggests similarity to our dataset, the scope in the SemEval dataset is much broader and includes any kind of production, e.g. blisters caused by a herpes virus, children “produced” by their parents, or questions asked by journalists.

The ACE guidelines for English relations (Linguistic Data Consortium, 2005) describe the relation type Agent-Artifact, which “applies when an agent owns an artifact, has possession of an artifact, uses an artifact, or caused an artifact to come into being.” However, the non-organization argument of the corresponding relation definition only allows facility-type entities, and does not mention products.

3. Bootstrapping Product Annotation

This section presents the data sources used in this work, and the pattern-based bootstrapping approach used for pre-annotating products.

3.1. Source Datasets

We collected a large dataset of web pages from business news portals, company home pages, and special interest forums, as well as posts from social media channels such as Twitter and Facebook. Web pages and forums were crawled based on an iteratively refined set of source URLs and keywords, such as company name lists. Similar keyword sets were also used for retrieving public Twitter and Facebook posts using the respective APIs of these services. The dataset was crawled over a period of 1.5 years, between July 2016 and December 2017, and consists of approximately 5.8 million documents. We focused on B2B information related sites, but did not explicitly exclude documents reporting on consumer products. Figure 1 shows an example document from the corpus.

As can be expected, documents exhibit a large degree of linguistic variance, ranging from journalistic writing to colloquially formulated tweets. In addition, HTML-to-text conversion and boilerplate removal are far from perfect, resulting in extraneous and not-well-formed content. We noticed that B2B news sites often appended “canned” company summaries at the end of news articles, which typically contain a lot of useful information about the company’s products. We included an example of this feature as the second paragraph of Figure 1.

3.2. Pattern-based Product Pre-Annotation

To gain a better understanding of the linguistic properties of the different types of product mentions in our dataset, we included an automatic pre-annotation step in our corpus analysis. Pre-annotation can help to decrease the duration of manual annotation and generally ease the annotation process (Kwon et al., 2014). We first developed a set of lexical patterns for the relation CompanyProvidesProduct. We chose this approach since we are also interested in identifying instances of this relation, and since it significantly increases the precision of product mention identification (at the cost of recall).

The CompanyProvidesProduct relation maps a company (organization) to products created, manufactured, provided, distributed or vendored by this company (Section 5.). Table 1 lists some example patterns. Many of these patterns can be varied by exchanging the verb or verb nominalization used, e.g. produce, create, develop, make, manufacture, offer. In total, we defined 13 base patterns, yielding a total of 173 surface patterns. We used a chunk parser to label potential product mentions. It matches noun phrases, optionally preceded or followed by adjectives or cardinal numbers, e.g. “high-resolution waveform analysis”, “High-Frequency 600mA DC-DC Buck Converter”, and “1500 ECL-PTU-208”. We included the VBG tag as in rare cases, gerund verb forms may be part of the product mention, e.g. “communicating sensors”. For pre-annotation, we randomly selected a set of 1,200 documents from the source dataset. Each document’s text was tokenized and part-of-speech tagged. Organizations were detected using Stanford NER (Manning et al., 2014).
Table 1: Example bootstrap patterns for the relation CompanyProvidesProduct used for pre-annotating product mentions. Company arguments are underlined, product arguments enclosed in brackets. For brevity, the chunking pattern applied to extract potential product mentions is only shown in the first row of the table.

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>ORG’s PRO: {&lt;VBG</td>
<td>NN.*</td>
</tr>
<tr>
<td>ORG [to produce</td>
<td>to manufacture</td>
</tr>
<tr>
<td>ORG [to be] [producer of</td>
<td>maker of</td>
</tr>
<tr>
<td>ORG [to be] [a</td>
<td>the</td>
</tr>
</tbody>
</table>

We then applied the patterns, which resulted in a total of 1,308 potential product mention matches. While many of these matches did at least cover a product mention, many results were unsatisfactory because they included excess lexical material, as shown in the following examples (square brackets denote the correct mention extent):

1. *highly accurate [3D magnetic sensor]*
2. *advanced [magnetic-hydraulic circuit breakers]*

We also observed that even when syntactical extraction worked correctly in a way that only phrases containing a product mention were retrieved, there were differences in semantic quality that should be taken into account. An adjective in the same position, for example, can either be of no value to the categorial specification of the product, in which case it can be neglected, or it can be a crucial part of the category description:

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. advanced [sensors]</td>
<td><em>advanced [sensors]</em> and <em>controls</em> as <em>products</em> in the sentence “Sensata Technologies develops sensors and controls.” during this second step.³ In a last step, coreferential mentions of these additional product mentions were also labeled in the remainder of the document, even if they did not occur as an argument of a relation mention. The reasoning here is that an NER algorithm should encounter consistent labels for the same token sequence, e.g. if “smartphones” is labeled as a product once, it should be labeled as a product everywhere in the document. All other product or product class references, i.e. those that were not part of a CompanyProvidesProduct relation mention, were not annotated as product entity mentions. We chose this strategy to limit the annotation effort for the initial corpus. In addition, disagreement by the annotators was very high when annotating all noun phrases that could potentially be viewed as products.</td>
</tr>
<tr>
<td>b. [magnetic sensors]</td>
<td>The extent to which these issues occurred seems to be closely linked with the specific product domain. These observations show that due to high boundary ambiguity as well as a broad syntactic and semantic variety of the surface variants of product mentions, it is necessary to define in advance which elements should be considered part of the extent of a product mention, and which should be excluded.</td>
</tr>
</tbody>
</table>

4. Annotation Guidelines

To formalize the annotation of product mentions and CompanyProvidesProduct relation mentions, we developed a set of annotations guidelines. For entity annotation, we base our guidelines on existing work, such as the ACE annotation guidelines (Doddington et al., 2004) for labeling organizations and companies. We try to follow similar guidelines for the annotation of products, but transform and expand these as detailed below.

Since many phrases in a typical document can be viewed as products or product classes (e.g. “mobile services”, “hotel chains”, “personal devices”), annotators limited their effort by adopting the following overall strategy for labeling a document: First, they annotated all name mentions of organizations and products, as well as any coreferential nominal and pronominal mentions of these. Coreference information was added as an extra relation type Identity. A single Identity relation was created for each coreference chain, with a source argument for the most precise name mention of an entity in a document, and target arguments for all other mentions of this entity. In a second step, annotators searched for occurrences of CompanyProvidesProduct relation mentions, and labeled (pro-) nominal product or product class references if they served as the argument of the identified relation mention. For example, annotators would label “sensors” and “controls” as products in the sentence “Sensata Technologies develops sensors and controls.” during this second step.¹ In a last step, coreferential mentions of these additional product mentions were also labeled in the remainder of the document, even if they did not occur as an argument of a relation mention. The reasoning here is that an NER algorithm should encounter consistent labels for the same token sequence, e.g. if “smartphones” is labeled as a product once, it should be labeled as a product everywhere in the document. All other product or product class references, i.e. those that were not part of a CompanyProvidesProduct relation mention, were not annotated as product entity mentions. We chose this strategy to limit the annotation effort for the initial corpus. In addition, disagreement by the annotators was very high when annotating all noun phrases that could potentially be viewed as products.

²We define as a product any commercially available good, be it a finished product, a pre-product, or a part or component of a larger product. While the focus of this work is on non-consumer products, this definition also includes consumer products. A product does not have to be a tangible object, but can be a service or virtual object. Although they are semantically closely related, product-like entity mentions that refer to an industry sector or branch of business are not treated as products. The industry term can, however, be part of the product mention. Categories such as brand names and trademarks also often appear as part of the product mention. For the reliable extraction of a product mention its maximum extent must be pre-defined. This means that one has to identify both the elements included in the extent and those outside of the extent that often appear alongside the product. |

1Only if these nouns were not labeled as coreferential to some product name mention in the first step, of course.
<table>
<thead>
<tr>
<th>Category</th>
<th>Example</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>company name</td>
<td>[Dunlop] Sport M3 winters</td>
<td>This is the name of the company that provides the product when it is mentioned as part of the product name. It is usually found in the beginning of the product mention. It tends to be a proper noun that does not always follow orthographic rules and sometimes appears as an abbreviation (e.g. “Tumblr”, “BMW”). The complexity of a company name tends to be reduced to the main word(s) when it is part of the product mention, i.e., “Toyota” instead of “Toyota Motor Corporation”.</td>
</tr>
<tr>
<td>brand name</td>
<td>Apple [iPhone] 6S</td>
<td>The brand name is the name under which a certain product is marketed. Usually this is a proper name, however it is not always capitalized or otherwise orthographically correct. Like all categories that can be realized as a proper name, a brand name can include any word class, special characters and punctuation (e.g. “FILL OR BUST!”).</td>
</tr>
<tr>
<td>series</td>
<td>VW Golf [VII]</td>
<td>This is the part of the product mention that denotes the series, generation, edition or model range. It is often realized as a number, sometimes as a name. Sometimes it includes the word “series” or “generation” or an equivalent abbreviation. Whether or not this can be part of a product mention is dependent on the domain.</td>
</tr>
<tr>
<td>model</td>
<td>BMW [i8]</td>
<td>This part of the product mention denotes the specific model of a product in a product series. It often consists of letters and numbers or a combination of the two.</td>
</tr>
<tr>
<td>trademark symbol</td>
<td>McRib[®]</td>
<td>The trademark symbol usually appears right after the brand name.</td>
</tr>
<tr>
<td>type</td>
<td>Nike Air Max 2016 [running [shoes]]</td>
<td>The type of the product is the broader category or subcategory a product falls into. The type is usually a common noun and can often be found at the end of the product mention. The category can include an attribute that serves as a specifier to the noun and it is the category in which nonspecific terms such as “product” or “solutions” can be included, serving as the head of the product mention, but only if specified further (e.g. “cosmetic product”).</td>
</tr>
<tr>
<td>feature</td>
<td>[2006] Ford Mustang [GT] Convertible [2-Door] Samsung Galaxy S7 [32 GB] [black]</td>
<td>Most other relevant aspects of products we will categorize as a feature. Since this is the vaguest category and the one that is most highly domain-dependent, it can be represented by a broad variety of linguistic manifestations. It includes elements such as the year of fabrication, colors, sizes, variants, and special features. Features can appear in almost any position in the product mention.</td>
</tr>
</tbody>
</table>

Table 2: Categories of product mention elements. Square brackets denote the extent of elements.

product mention. We will discuss the different elements a product mention can consist of, considering both semantic categories and their word class counterparts, as well as elements that are excluded from the product extent according to our annotation schema.

The ways in which a product can be mentioned in a text are manifold:

(3)  
  a. vehicle  
  b. SUV  
  c. Land Cruiser  
  d. Toyota Land Cruiser  
  e. Toyota Land Cruiser 100 Series VX  
  f. Toyota Land Cruiser 100 Series VX SUV

All of these examples are possible ways to refer to the same real-world product and could appear as the product argument in a relation expressing a product the company “Toyota” sells. 3[a.] and 3[b.] are rather vague, describing a product category, 3[c.] and 3[d.] are more specific, distinguishing the car from all other brands by all other companies, and 3[e.] and 3[f.] are so specific that the product cannot be confused with another model.

Product mentions are generally realized as noun phrases, containing at least one proper noun or one common noun. As a proper noun, the head of the noun phrase can consist of individual letters or numbers or a series of numbers and/or letters:

(4)  
  a. AP3405  
  b. 1500 ECL-PTU-208  
  c. Samsung 14nm LPP Process

Often, the noun is accompanied by further distinctive attributes that can appear in different word classes as illustrated in the following examples:

(5)  
  a. smart sensors (adjective)  
  b. communicating sensors (verb, gerund)  
  c. vision sensors (common noun)  
  d. Hall sensors (proper noun)

### 4.2. Elements of Product Mentions

We found a limited set of elements that products usually consist of. This set can be subdivided into seven categories: company name, brand name, series, model, trademark symbol, type and feature. Not all of these elements appear in every product mention. Product mentions can vary strongly in length and complexity, from a single element (6[a.-c.]), to a combination of any of the categories (6[d.-f.]) to a coverage of all of the categories (6[g.]).
a. sensors (type)
b. Kleenex (brand name)
c. Q7 (model)
d. Audi Q7 (company name and model)
e. Innocent Drinks smoothies (company name and type)
f. white iPhone 6 (feature, brand name and model)
g. Toyota Land Cruiser 100 Series VX SUV diesel turbo (all of the above)

As all of the examples used thus far have shown, some categories are more essential to a product mention than others. A product mention contains at least a common noun, representing the product type, or a proper noun that can either refer to a brand name or a specific model. Like brand names and models, the company name that often appears as part of the product mention (but is not essential to it) can also consist of a proper name and therefore include any kind of word class and even punctuation. While most of the seven categories can include or can be realized as nouns, series and generations as well as models tend to consist of letters or numbers or combinations of the two. Table 2 lists the further specification of the individual categories as well as examples for each category. In the table, the different categories of elements that we consider part of product mentions are listed in the order in which they usually appear when a product mention contains more than one category, with the exception of the category feature that can be found in any position.

Since it is not always apparent which category a part of a product mention falls into – sometimes the same part of a product mention could be assigned to two or more categories – the nested elements that constitute a product mention are currently not annotated, but only used by the annotator to determine the product mention’s extent.

### 4.3. Elements Excluded from Product Mentions

One of the major results of our analysis is that there are a number of elements that often appear alongside a product mention and may be mistaken as part of it. Unless they are included in the proper name of the product (usually the company name or the brand name part), articles, prepositional phrases and prepositions, relative clauses and appositions are never considered part of the product extent unless they serve to define the product further. If they do, but there is more than one attribute fulfilling that function, we have to differentiate between products that are described by several attributes on the one hand and different product mentions that share a head but are distinguished by the attributes on the other. In the former case, the commas or linking conjunctions are included in the extent of the product, in the latter case they are not, but two – or more – product mentions are annotated. The following examples illustrate this issue:

### 5. The CompanyProvidesProduct Relation

The CompanyProvidesProduct relation consists of two mandatory arguments, a company (organization) and a product, as well as of one optional argument, a trigger. A company can serve as the first argument if it is stated as the creator, manufacturer, provider, distributor or vendor of the product argument. The slot for the second mandatory argument can be filled by one or more product mentions (e.g. in the case of conjunctive enumerations). Trigger concepts are a generic class of annotations that cover lexical expressions (terms or phrases) or syntactical elements (e.g. possessive marker -s or prepositional constructs) that indicate a specific event type.

The annotators were instructed to annotate only relation instances mentioned within a sentence. The following examples illustrate the relation annotation:

a. [Parkifi]company is a fast-growing technology company focused on [providing]trigger their customers with [real-time parking data]product

c. [BMW]company [s]trigger [Z3]product
d. [Intuition Executive]product [by]trigger [Honeywell]company collects and analyzes large amounts of data
e. [[Apple]company[Watch Series 2]]product

Our annotation guidelines also consider some specific cases. For example, if a sentence contains a full-length company name followed and coreferenced by the company
abbreviation, then we label both mentions as individual company mentions, but only a single relation mention, between the full-length company mention and the product, is annotated. The relation between the company acronym and a product is implicitly given by the coreference information. Example 9 illustrates this issue: the company name IS International Services LLC, the trigger providing and the product engineering services are annotated as relation arguments, while the company’s abbreviation IS are connected by the Identity relation.

\begin{equation}
\text{a. [IS International Services LLC]}_{\text{company}} \rightarrow \text{[IS]}_{\text{company}} \text{ is a uniquely qualified business [providing]}_{\text{trigger}} \text{[engineering services]}_{\text{product}}
\end{equation}

Furthermore, if a sentence contains more than one trigger for the same relation instance, then as many relation mentions are annotated as there are triggers. Example 10 contains one company, one product, and the three triggers developer, manufacturer and vendor, each of them referring to a different way of how the product is related to the company – therefore three relation mentions are created.

\begin{equation}
\text{b. FUJIFILM invested in [Japan Biomedical Co.]}_{\text{company}} \rightarrow \text{[developer]}_{\text{trigger}} \text{[manufacturer]}_{\text{trigger}} \text{and [vendor]}_{\text{trigger}} \text{of [additives for cell culture media]}_{\text{products}}
\end{equation}

6. Corpus Statistics

<table>
<thead>
<tr>
<th>Total</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td># Documents</td>
<td>152</td>
</tr>
<tr>
<td># Sentences</td>
<td>4001</td>
</tr>
<tr>
<td># Words</td>
<td>131929</td>
</tr>
<tr>
<td># Companies</td>
<td>2191</td>
</tr>
<tr>
<td># Products</td>
<td>1717</td>
</tr>
<tr>
<td># CompaniesProvidesProduct</td>
<td>379</td>
</tr>
</tbody>
</table>

Table 3: Corpus Statistics

This section describes the corpus of documents annotated with product mentions, including product parts, technologies, and product classes, using the guidelines described in the previous section. Documents included in the corpus are sampled from the dataset that we described in Section 3. Table 3 lists some statistics of the current state of the corpus. The annotation is being carried out by two trained linguistics students. In cases of disagreement, a third expert annotator is consulted to reach a final decision. The current datasets consists of 152 documents with more than 131,000 words. Thus far, 2, 908 entity mentions (2, 191 organizations, 1, 717 products) have been annotated, and a total of 379 CompanyProvidesProduct relation mentions. For the corpus annotation we use the markup tool Recon (Li et al., 2012), which allows annotating n-ary relations among text elements. Recon provides a graphical user interface that enables users to mark arbitrary text spans as entities, to connect entities to create relations, and to assign semantic roles to argument entities. Since the corpus is still in the process of being created, we cannot report any reliable inter-annotator agreement scores at the moment. We will include information about inter-annotator agreement at the entity and relation mention level in the final release. The corpus and the guidelines will be made available at https://dfki-lt-re-group.bitbucket.io/product-corpus. We distribute the dataset in an AVRO-based compact binary format, along with the corresponding schema and reader tools.

7. Conclusion

In this work we presented a fine-grained analysis and annotation schema for mentions of product entities and CompanyProvidesProduct relations in English web and social media texts. The schema is motivated by linguistic aspects and addresses the needs of recognizing industry- and product-related facts and relations. We presented a semi-automatic annotation process in order to ease the annotation procedure. While we have only annotated a small set of documents so far, the annotation effort to increase the size of the corpus is ongoing.

Acknowledgments

This research was partially supported by the German Federal Ministry of Economics and Energy (BMWi) through the projects SD4M (01MD15007B) and SDW (01MD15010A) and by the German Federal Ministry of Education and Research (BMBF) through the project BBDC (01IS14013E).

8. Bibliographical References


