Chapter 5

Term Labeling

In the previous chapters we have discussed approaches to ATR and variation recognition. The outputs of these processes are terms ranked on the basis of their termhood scores. At this point, we do not know whether a term such as *tuberculose* is a name of a *disease, treatment*, or *virus*. These labels can only be assigned in the next step, that is term labeling.

In general, the task of term labeling is to disambiguate the senses of a term. Term labeling may help to map a term according to its categories in an ontology or thesaurus, or to understand the roles of a term in a relation (e.g., as actors, sources, objects) (Hatzivassiloglou et al., 2001; Gaizauskas et al., 2003; Rice et al., 2005; Spasic et al., 2003; Koike et al., 2005). For a medical domain, the most comprehensive and widely used thesaurus is the UMLS. It maps medical terms of many languages into 135 semantic types. However, most of the labeled terms are in English.

This chapter aims to answer research question #5: to what extent terms and their labels in the UMLS can be used to classify unseen terms in a non-English language, such as Dutch? Unlike previous labeling methods (Nobata et al., 1999; Spasic et al., 2003; Rice et al., 2005), which typically rely on machine learning techniques, our approach is based on a heuristic that utilizes the UMLS Metathesaurus entries.

We describe previous work on term labeling in Section 5.2, and then a brief description on the UMLS and on how we pre-process it in Section 5.3. In Section 5.4 we explain our term labeling method, which is followed by experiments in Section 5.5. We evaluate the results and compare the precision with a previously reported experiment on the same data source in Section 5.6. Section 5.7 summarizes our method and evaluation presented in this chapter.

5.1 Introduction

If we want to build a medical QA system, a set of medically relevant relations should be extracted. Medical relations can be expressed in text by rather general linguistic patterns, such as *X may lead to Y* or *X occurs in Y*. Such patterns can nevertheless be used to extract medical relations with high accuracy if we require that both X and Y are medical terms. We may also impose the restriction that X and Y have to be terms that belong to a given class (i.e. X and Y are medical terms denoting, respectively, a *Virus* and a *Disease*).

Our term labeling task is aimed at classifying medically relevant terms contained in these kinds of relations. The assigned classes will further improve the accuracy of the relation extraction.

As for the medical classes, we decided to adopt labels used in the UMLS, because it is one of the most comprehensive ontological resources for the medical domain. In UMLS, all medical terms collected from various terminological sources have been labeled with a set of common semantic types. A large amount of effort has been spent in building and maintaining this ontology and in providing accurate labeling.

We use the UMLS Metathesaurus entries, which contain terms and their semantic types, to recognize new terms by matching their strings or root forms. The labels for the new terms are taken from the most frequent label(s) used by the matching entries. Since we use the UMLS as our labeling resource, a problem related to the availability of labels for Dutch terms in the UMLS occurs. This resource contains only 163,000 labeled terms for Dutch compared to almost 3 million labeled terms for English. To increase coverage, when for example the term is in Dutch and the UMLS entry is in English, we use machine translation to find a match.

5.2 Previous Work

Typically, term classification tasks rely on machine learning techniques. They are mostly based on statistics, such as Hidden Markov Models (HMM) and naive Bayes, and often use decision trees, rule induction, Support Vector Machines (SVM), and genetic algorithms.

Nobata et al. (1999) compare two classification methods based on statistics and decision trees in a task of classifying terms from MEDLINE abstracts. In the first method, they classify terms by computing the similarity of the terms to the distribution of words in a pre-classified word list from databases. Since the word list is rarely complete, it can be extended with the output of word clustering. In the second method, they use several feature sets including PoS, morphology, and a list of words specific to the domain. They found that the statistical method is comparatively better at classifying DNA and RNA, while the decision tree method is better at classifying *Source* and *Protein* classes.

Since both methods depend on a closed list of words, they are less suitable for unseen words. To overcome this problem, Collier et al. (2000) use richer word features such as DigitNumber, SingleCap, GreekLetter, CapsAndDigits, TwoCaps, Hyphen, Backslash, Colon, Percent, etc. They believe that such features can model the similarities between known words in the training data and unknown words in the test data. For example, a new word *AP-1* can be classified to a target class because there exists a known word *LMP-1* which shares some similar features (TwoCaps and Hyphen) with it. Compared to PoS information, these features are more meaningful, since PoS will predominantly be noun for all names. The results of their HMM-based method show that the character features add 10.6% to the F-score. However, this method suffers from data sparseness. For example, proteins get the best result compared to RNA, since most of the classes in the training data are proteins.

While the above methods are based on internal evidence of terms, Rice et al. (2005) use the co-occurrence of terms with a protein of interest as a base to classify the role of unseen proteins that have similar co-occurrences of terms. They use a method based on SVMs to learn the relevant and informative co-occurring terms. The method works well as the number of relevant documents to a particular protein increases, otherwise it works poorly.

Spasic et al. (2003) use a genetic algorithm as a learning engine for the classification task. Their approach is based on verb complementation patterns automatically learned by combining information found in a corpus and in an ontology (UMLS) for the biomedical domain. The extracted verbs from the corpus are ranked based on their frequency of occurrence in the corpus and based on their co-occurrence with terms. These terms can be seen as parameters of their collocated verbs. The score of each class for a term is calculated using a genetic algorithm containing a parameter which balances the impact of the class probabilities and the similarity measure. The highest class score is used to classify the term, or alternatively multiple classes are assigned if there is no highest class. This method gets 63.83% average precision and 12.20% recall (20.48% F-measure).

The current approaches to term classification described above are mainly based on learning algorithms. In order to achieve high performance, they require a large training set containing relevant classes, otherwise these methods will suffer from data sparseness. Obviously, we will run into the same problem if using the same approaches, since our data set is small.

We are inspired by Nobata et al. (1999) who use a set of pre-classified words from databases to classify terms from text. For the medical domain, this idea is worth an attempt as there is a large database or thesaurus i.e. UMLS. The UMLS Metathesaurus contains a large number of pre-classified terms (2.10 million terms or 4.7 million term-class labels). This will answer the problem of data sparseness when using a machine learning technique. However, UMLS is not a kind of training dataset that can be used directly for such a learning algorithm. We hypothesize that with an appropriate method, the UMLS Metathesaurus can be used to classify unseen terms.

Another challenge we face in our term classification task is a multilinguality. We work on Dutch terms while UMLS is mainly in English. We need to solve this language barrier so the classification task can use most of the UMLS term entries.

5.3 Resources

5.3.1 The UMLS Metathesaurus

The UMLS Metathesaurus[1] is a multi-lingual vocabulary containing concepts related to biomedical and health built from various “source vocabularies” such as thesauri, classifications, code sets, and controlled terms. It also contains the variation names of the concepts and the relationships among them.

In the UMLS, the entire concept structure is presented in a single file, i.e. `MRCONSO.RRF`, and organized by concept or meaning. For example, Table 5.1 shows a concept with the identifier `C0000039`, its language, its sources, and its variation names. All of the variation names in this table refer to the same

[1] http://www.nlm.nih.gov/research/umls/about_umls.html#Metathesaurus

Identifier	**Lang**	**Source**	**String**
C0000039	ENG	D015060	1,2-Dipalmitoylphosphatidylcholine
C0000039	ENG	C25778	1,2-Dipalmitoylphosphatidylcholine
C0000039	ENG	NOCODE	1,2-Dipalmitoylphosphatidylcholine
C0000039	ENG	D015060	1,2 Dipalmitoylphosphatidylcholine
C0000039	ENG	D015060	1,2-Dihexadecyl-sn-Glycerophosphocholine
C0000039	ENG	D015060	1,2 Dihexadecyl sn Glycerophosphocholine
C0000039	ENG	D015060	1,2-Dipalmitoyl-Glycerophosphocholine
C0000039	ENG	F-63675	Dipalmitoylphosphatidylcholine
C0000039	ENG	MTHU010538	Dipalmitoylphosphatidylcholine

Table 5.1: A concept from the `MRCONSO.RRF` file with its identifier, language, sources, and variation names.

```
C0000039|T119|A1.4.1.2.1.9|Lipid|AT17617573||
C0000039|T121|A1.4.1.1.1|Pharmacologic Substance|AT17567371||
```

Figure 5.1: Examples of entries in the file `MRSTY.RFF` that relate concepts (identifiers) to their semantic types.

concept (represented by the same identifier) and apparently are in the same language. Some vocabulary sources use the same string for this concept, for example, the string *1,2-Dipalmitoylphosphatidylcholine* is used in the sources `D015060` and `C25778`. Other sources, such as C25778 and F-63675, use different strings (*1,2-Dipalmitoylphosphatidylcholine* and *Dipalmitoylphosphatidylcholine*) that lead to variations.

The UMLS version *2007AB July 2007* we are using for this thesis contains 4.7 million entries in its `MRCONSO.RRF` file, of which 4.5 million are in English and only 216.000 are in Dutch. There is no entry from other languages since we only exported entries of these two languages into the file. The number of unique concepts in the file is 1.4 million, and thus on average each concept has 3 or 4 variation names.

Each of the concepts in the UMLS Metathesaurus is categorized by at least one semantic type from the UMLS Semantic Network described in the next subsection (5.3.2). This categorization, which is presented in the file `MRSTY.RFF` of the Metathesaurus, gives a consistent classification to the concepts at general levels. Consider, for example, the concept with the identifier `C0000039` above. This concept is categorized to two semantic types, i.e. *Lipid* and *Pharmacologic Substance*, as shown by Figure 5.1. Each line in the file contains a concept identifier, a semantic type identifier, a semantic type tree number, a semantic type, and an attribute identifier.

Using these two files, the meaning of each term in the Metathesaurus can be derived by looking at several related information features, such as its source, context (the position in the semantic type tree), variation names, and synonyms. For the current task, the files are used to get a set of <*instance, class*> relations or term-to-class relations. How we extract and pre-process this information is described in subsection 5.3.3.

5.3.2 The UMLS Semantic Network

The UMLS Semantic Network is a set of subject categories (Semantic Types) and relationships (Semantic Relations) that is used to classify and relate the Metathesaurus entries. The purpose of this network is to provide a consistent mapping between terms and their classes and among terms themselves, across various terminology systems.

The Semantic Network contains 135 semantic types and 54 semantic relationships. In this Network, each semantic type can be seen as a node and each semantic relationship as an edge that links the nodes. Consider, for example, a portion of the Semantic Network in Figure 5.2(a). It shows a hierarchical structure of a group of semantic types starting from the highest node *Physiologic Function* to the lower nodes *Mental Process* and *Genetic Function.* Each lower node is linked to its parent node by an "isa" relation which is one of the relations in the set of Semantic Relations. This group is actually within a larger group which is under the *Event* type, as shown in Figure 5.2(b).

We use all of the classes in the UMLS Semantic Types for our term labeling task. And since we classify terms to classes, only the "isa" relation of the UMLS Semantic Relations, that ties terms to classes, will be considered.

5.3.3 Indexing Terms and Their Semantic Types

The UMLS provides a tool for accessing the concepts and relationships, namely the MetamorphoSys. We use this tool to select "source vocabularies" that contain terms in Dutch and English, and then export the entries into the `MRCONSO.RRF` and `MRSTY.RRF` files. Having the installed knowledge sources, we can use the tool to search a concept and its relationships, or use library programs provided by the UMLS to programmatically access the database.

We decided to use an information retrieval tool which provides simple and fast accesses to its index and returns customized output based on the information we need. The benefit of using such a tool is that we can apply various matching techniques, for example, exact matches, stemming, and boolean matches. These features are important for our term labeling strategy since we attempt to find terms in the database that are similar to unseen terms from text. For this purpose, we chose Solr,[2] an open source search server based on the Lucene Java[3] search library. One of the most important features is its ability to index and search documents based on our customized indexing rules.

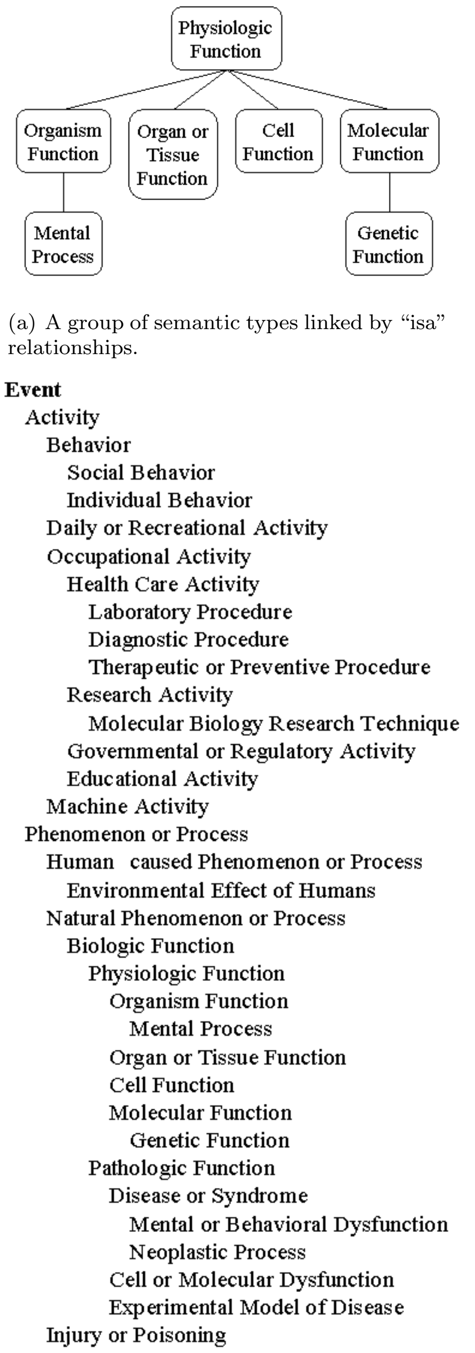
Inputs to Solr are documents in a standard XML format. Each Solr document contains an identifier field and one or more document specific fields. Each field has a `name` attribute that defines which rule (or analyzer) should be used during the indexing and searching and which field should be searched during processing queries.

To build our Solr index, we extract all variation names as well as their semantic types from the installed knowledge sources. Each of our Solr documents represents a variation name that has 9 fields as shown in Table 5.2. This table shows the attributes of these fields, their descriptions, and their types.

Most of the fields are analyzed as *string*s, in which text is indexed and stored verbatim (without tokenization and stemming). Unlike the *string* type, the *text*

[2] Solr, http://lucene.apache.org/solr/

[3] Lucene, http://lucene.apache.org/java



(a) A group of semantic types linked by "isa" relationships.

- **Event**
  - Activity
    - Behavior
      - Social Behavior
      - Individual Behavior
    - Daily or Recreational Activity
    - Occupational Activity
      - Health Care Activity
        - Laboratory Procedure
        - Diagnostic Procedure
        - Therapeutic or Preventive Procedure
      - Research Activity
        - Molecular Biology Research Technique
      - Governmental or Regulatory Activity
      - Educational Activity
    - Machine Activity
  - Phenomenon or Process
    - Human caused Phenomenon or Process
      - Environmental Effect of Humans
    - Natural Phenomenon or Process
      - Biologic Function
        - Physiologic Function
          - Organism Function
            - Mental Process
          - Organ or Tissue Function
          - Cell Function
          - Molecular Function
            - Genetic Function
        - Pathologic Function
          - Disease or Syndrome
            - Mental or Behavioral Dysfunction
            - Neoplastic Process
          - Cell or Molecular Dysfunction
          - Experimental Model of Disease
    - Injury or Poisoning

(b) The hierarchy of semantic types under the *Event* type.

Figure 5.2: A portion of the Semantic Network taken from the UMLS Documentation, http://www.nlm.nih.gov/research/umls/meta3.html (11 February 2008).

Field	**Description**	**Type**
id	unique document ID	string
cui	unique concept ID	string
stn	semantic tree number	string
sty	semantic type	string
lang	language	string
src	source	string
hd	head words	string
root	root words	*string_tok*
term	term strings	text

Table 5.2: Field names and their analyzer for each UMLS term. The *string_tok* is our customized type.

type applies a tokenization, a lower casing, and a porter stemming filter. Both of these types are provided by Solr. For the `root` field, we create a new type which is based on the *string* type, namely the *string_tok* type. Besides applying a whole-string matching, this type also applies a lower casing and a porter stemming filter during Solr's indexing and searching processes. These filters are required for the English terms since, unlike the Dutch terms, we do not parse the English terms to get their `head`s and `root`s.

Figure 5.3 gives two examples of Solr documents, each representing a term in Dutch and in English, respectively. The first document is about the *foetale misvorming* 'fetal malformation' term that after being parsed has *foetaal misvorming* as its root and *misvorming* as its head. This term is from the *Medical Dictionary for Regulatory Activities Terminology (MedDRA), Dutch Edition, 7.0*, which is abbreviated as *MDRDUT*, and is classified to the *Congenital Abnormality* type in the UMLS. The second document is about the *congenital anomalies* term. Since it is not parsed by the Alpino parser, its `root` field is simply a verbatim copy of its `term` field, and its `head` field is not provided. However, during the indexing and searching processes, Solr will stem this string using the porter stemming filter. This term is from the *International Classification of Primary Care, Version 2-Plus, 2000*, which is abbreviated as *ICPC2P*, and has the same label as the first document.

The effect of assigning the *string_tok* type to the `root` field can be illustrated by the second document. Let assume that a query term *aangeboren afwijking*, translated into English as *congenital anomaly*, is an unseen term from our Dutch text. Having the translation, we search on the `root` field for English terms overlapped with the translation. And since *congenital anomalies* and *congenital anomaly* have the same stem, Solr will return the second document as a matching result.

Our Solr index contains 3,142,578 terms, of which 95% are in English and only 5% are in Dutch. These numbers clearly show that English terms are the largest portion in the UMLS and worth of exploitation for classifying terms of other languages.

```
<doc>
  <field name="id">C0000768-1372</field>
  <field name="cui">C0000768</field>
  <field name="stn">A1.2.2.1</field>
  <field name="sty">Congenital Abnormality</field>
  <field name="lang">DUT</field>
  <field name="src">MDRDUT</field>
  <field name="hd">misvorming</field>
  <field name="root">foetaal misvorming</field>
  <field name="term">foetale misvorming</field>
</doc>
<doc>
  <field name="id">C0000768-1308</field>
  <field name="cui">C0000768</field>
  <field name="stn">A1.2.2.1</field>
  <field name="sty">Congenital Abnormality</field>
  <field name="lang">ENG</field>
  <field name="src">ICPC2P</field>
  <field name="hd"></field>
  <field name="root">congenital anomalies</field>
  <field name="term">congenital anomalies</field>
</doc>
```

Figure 5.3: Examples of Solr documents for Dutch and English terms.

5.3.4 Corpus

To evaluate our term labeling method, we use the IMIX medical corpus[4] that consists of text from a medical encyclopedia and a medical handbook. This corpus has been manually annotated with relations and concepts. An example of the annotation is given in Figure 5.4 where a sentence is tagged with a *definition* relation and several terms are tagged with *disease* (i.e. *RSI*) and *body_part* concept types (e.g. *boven rug* 'upper back', *nek* 'neck', and *schoudergebied* 'shoulder area').

This annotation is useful for evaluating the accuracy of our labeling method. Compared to the current release of the UMLS Semantic Network that contains 135 semantic types, this corpus was only labeled with 11 semantic types, namely: *bodily_function*, *body_part*, *disease*, *disease_feature*, *disease_symptom*, *duration*, *method_of_diagnosis*, *microorganism*, *person*, *person_feature*, and *treatment*. To evaluate our labeling results, which based on the UMLS semantic types, we map the most frequently used UMLS labels to corpus labels. The complete mapping is presented in Table 5.3.

5.3.5 Dictionary

We use the Google Translation[5] service to translate terms from Dutch to English. To achieve a high performance in terms of the processing speed, we create

[4] Developed in the Tilburg University IMIX/Rolaquad project. http://ilk.uvt.nl/rolaquad

[5] Google Translation, http://google.com/translate_t?langpair=nl|en

```
<rel_definition_of id="126">
 <con_disease id="27">RSI</con_disease> is een verzamelnaam
 voor klachten, symptomen en syndromen die voorkomen in
 <con_body_part id="112">bovenrug</con_body_part>,
 <con_body_part id="87">nek</con_body_part>- en
 <con_body_part id="99">schoudergebied</con_body_part>,
 <con_body_part id="102">armen</con_body_part>,
 <con_body_part id="111">ellebogen</con_body_part>,
 <con_body_part id="86">polsen</con_body_part>,
 <con_body_part id="97">handen</con_body_part> en
 <con_body_part id="107">vingers</con_body_part>.
</rel_definition_of>
```

Figure 5.4: An example of annotation in the IMIX medical corpus for the sentence *RSI is een verzamelnaam voor klachten, symptomen en syndromen die voorkomen in bovenrug, nek, schoudergebied, armen, ellebogen, polsen, handen en vingers* 'RSI is a collective term for complaints, symptoms and syndromes that occur in upper back, neck, shoulder area, arms, elbows, wrists, hands and fingers'.

Corpus label	**UMLS label**
body_part	A1.2.3.4;A1.2.1;A1.2.1;A1.2.3.5;A1.4.1.2.1;A1.4.1.1.2;A1.4.1.2.1.9;A1.4.1.2.1.8;A1.2.3.2;A1.2.3.1;A1.2;A1.2.3.3;A2.1.4.1;A1.4.2;A2.1.5.1;A2.1.5.2;A1.4.1.1.3.2;A1.4.1.2.1.7;
disease	B2.2.1.2.1;B2.2.1.2;B2.2.1.2.1.2;B2.3;B2.2.1.2.1.1;A1.2.2.1;A1.2.2;
disease_symptom	A2.2.2;A2.2;B2.2.1.2;B2.2.1.2.1;
disease_feature	B1.1.2;A2.1.4;A2.1.2;B2.2.1.2;
person	A2.9.3;A2.9.4;A2.9.5;A2.9.2;A2.9.5;A2.9.1;A1.1.7.2.5.1;A1.2.1;
person_feature	A2.3;A2.2;B2.2.1.1.1.1;A2.1.3;A2.1.4;A2.1.1;B1.1.2;
microorganism	A1.1.5;A1.1.3;A1.1.2;A1.1.7.1;A1.1;A1.2.3.3;A1.2.3.4;A1.1.4;A1.4.1.2.1;
treatment	B1.3.1.3;B2.2;B1.3.1.1;A1.4.1.1.1;A1.4.1.1.2;A1.4.1.2.1;A1.4.1.1.1.1;A1.3.1;A1.4.1.2.1.9.1;B1.2;B1.3.1;A2.3.1;
bodily_function	B2.2.1.1;B2.2.1.1.1;B2.2.1.1.2;B2.2.1.1.1.1;B2.2;B1.1;B2.2.1.1.3;B2.2.1;A2.1.4;
method_of_diagnosis	B1.3.1.2;B1.3.1.1;B1.3.1.3;B1.3.1;
duration	A2.1.1;

Table 5.3: Mapping the UMLS labels to the corpus labels. For the UMLS labels, we use their semantic tree numbers. The readers can find the meaning of the UMLS labels from the SRDEF table of the UMLS Semantic Network (http://www.nlm.nih.gov/research/umls/meta3.html).

a table that maps each term from the corpus to its corresponding translation.

We use the machine translation engine rather than a bilingual dictionary since there are no good publicly accessible medical EN-NL dictionaries. Besides that, a large portion of our terms, especially those that occur less frequently in the corpus, are multi-word terms which often missing in the dictionary. Since Google Translation is based on n-grams, we have an intuition that the probability of getting translations from this engine is higher.

5.4 Method

Our experiment in Section 3.7 on *Using Multilingual Terminologies* for term extraction has concluded that "an existing multi-word terminology is useful for identifying new multi-word terms of a particular language, as long as there is overlap at the word level." The experiment suggested that there are two strategies to increase the overlap, namely stemming and translation. Using stemming, two words with different suffixes will overlap if they map to the same stem. While using translation, two words from different languages will overlap if one translates to the other.

In this section, we explore the translation approach to obtain an overlap or a match between two terms on different languages. Our purpose is to find the most accurate labels for new terms (in Dutch) using the overlap with terms in the other language (English).

Our method is motivated by the fact that medical terminology differs across languages, but also is closely related. Technical medical terms in Dutch, for instance, often are simply borrowed from English (i.e. *stress, borderliner, drugs* and acronyms like ADHD and PTSS), or are cognates (i.e. English *genetic* and Dutch *genetisch*). Some terms are genuinely different in the two languages (*infection* and *besmetting*), and need to be translated.

5.4.1 Example-based Labeling

Our labeling method can be best described as an example-based strategy. It begins with collecting samples similar to the query term, and then selects the best label from these samples. Since we have a large number of entries in the database, we can retrieve enough samples for the label.

Each entry in the database has a number of attributes, i.e. a term string, a root form, a head word, and words. These attributes are evaluated in a particular order to get the most similar samples. Some attributes, such as string and root form, are good indicators for similarity and should be evaluated first. For example, the sample *congenital anomalies* will match at the `string` level with the query term *congenital anomalies* and at the `root` level with the query term *congenitale anomalie.* If there is no sample with such a high similarity, then we search for less similar samples by evaluating less significant attributes, until a similar term is found.

Each sample has a 'type' attribute indicating the class it belong to. Often the values of this attribute among the samples are not the same, especially if the samples are less similar to the query term. Even, if the samples are highly similar to the query term, their types can be different depending on the sources and the context of the samples. Moreover, a sample may also have more than one type, for example, the term *nerve* is classified into two types: *Tissue* (1.2.3.2) and *Body Part, Organ, or Organ Component* (A1.2.3.1). To decide which type will be best assigned to the query term, we look at the labels assigned to the terms. The most frequent maximally 3 labels will be selected.

5.4.2 The Algorithm

This subsection formalizes the labeling method described in the previous subsection. To classify Dutch terms on the basis of a subset of UMLS concepts that contains Dutch and English terms, we use a sequence of five steps below.

Step 1: Exact match of root forms for Dutch term.
Query term: *psychish aandoening* (root forms)
Solr query: `?q=root:"psychish aandoening"+AND+lang:DUT`

Step 2: Exact match of term string for Dutch term.
Query term: *psychische aandoeningen* (term string)
Solr query: `?q=root:"psychische aandoeningen"+AND+lang:DUT`

Step 3: Exact match of translated term in English.
Query term: *mental disorders* (translation)
Solr query: `?q=root:"mental disorders"+AND+lang:ENG`

Step 4: Exact match of head word for Dutch term.
Query term: *aandoeningen* (head word)
Solr query: `?q=hd:"aandoeningen"+AND+lang:DUT`

Step 5: Match one of the words in the translated term. Since some terms do not have any translation, this query is applied to both languages.
Query term: *mental disorders* (translation)
Solr query: `?q=term:"mental"+OR+term:"disorders"`

If step 1 returns no result, step 2 will be evaluated, and so on. Consider, for example, the query term *psychische aandoeningen* ‘mental disorders’ used in the steps above. After several queries, a set of samples similar to the query term is returned at step 3, as shown in Table 5.4. There are two sample terms that match with that query term, i.e. *mental disorders* and *mental disorder*, and apparently both have the same label. And since there is only one label within the samples, namely *Mental or Behavioral Dysfunction* (B2.2.1.2.1.1), this label will be assigned to the query term.

Term	**Source**	**Semantic type**
mental disorders	SNMI	B2.2.1.2.1.1:Mental or Behavioral Dysfunction
mental disorder	SNOMEDCT	B2.2.1.2.1.1:Mental or Behavioral Dysfunction

Table 5.4: Terms in the UMLS that match with the query term *psychische aandoeningen* ‘mental disorders’ at step 3 of the heuristic.

In case there is more than one label (this happens especially at steps 4 and 5), a further heuristic is needed to select the best one. Assume that a query term *besmettelijk enterovirus* and its translation ‘infectious enterovirus’ are not found in the UMLS. In that case Step 1 until step 4 of the heuristic would

```
Solr query: ?q=term:"infectious"+OR+term:"enterovirus"
Solr response:
{
 "responseHeader":{
  "status":0,
  "QTime":3993,
  "params":{
        "wt":"json",
        "rows":"100",
        "start":"0",
        "indent":"on",
        "fl":"*,score",
        "q":"term:infectious OR term:enterovirus",
        "qt":"standard",
        "version":"2.2"}},
  "response":{"numFound":5542,"start":0,"maxScore":7.6562815,
        "docs":[
        {
         "id":"C0014378-59078",
         "sku":"C0014378-59078",
         "cui":"C0014378",
         "stn":"B2.2.1.2.1",
         "sty":"Disease or Syndrome",
         "lang":"DUT",
         "src":"MSHDUT",
         "hd":"",
         "root":"enterovirus- infectie",
         "term":"enterovirus- infectie",
         "timestamp":"2007-11-19T11:20:53.841Z",
         "popularity":0,
         "score":7.6562815},
        {
         "id":"C1400756-2471964",
         "sku":"C1400756-2471964",
         "cui":"C1400756",
         "stn":"B2.2.1.2.1",
         "sty":"Disease or Syndrome",
         "lang":"DUT",
         "src":"ICPC2ICD10DUT",
         "hd":"",
         "root":"enterovirus viraal; infectie",
         "term":"enterovirus viraal; infectie",
         "timestamp":"2007-11-19T11:39:32.272Z",
         "popularity":0,
         "score":6.1250253},
         ...
     ]}
}
```

Figure 5.5: An example of Solr query and results for the query term *infectious enterovirus*.

Term	Source	Lang	Semantic type
enterovirus	SNMI	ENG	A1.1.3
enteroviruses	NCBI	ENG	A1.1.3
enterovirus echo	SNM	ENG	A1.1.3
enterovirus; meningitis	ICPC2ICD10	DUT	B2.2.1.2.1
enterovirus diseases	LCH	ENG	B2.2.1.2.1
genus enterovirus	SNOMEDCT	ENG	A1.1.3
302 enterovirus	SNMI	ENG	A1.1.3
varkens- enterovirus	MSHDUT	DUT	A1.1.3
diseases due to enterovirus	SNMI	ENG	B2.2.1.2.1
enterovirus coxsackie b1	SNM	ENG	A1.1.3
enterovirus, centraal- zenuwstelsel infectie; viraal	ICPC2ICD10	DUT	B2.2.1.2.1
unspecified enterovirus infection	ICD10AM	ENG	B2.2.1.2.1

Table 5.5: Examples of terms and their semantic types that match with the query term *infectious enterovirus*.

not return any result for this query term. Finally, step 5 searches for all terms containing at least one of the words ‘infectious’ or ‘enterovirus’.

The result of a succeeding step is presented in Figure 5.5 using a JSON[6] (see the `"wt"` field) format, showing that Solr has found 5542 documents (see the `"numFound"` field) from its index, of which 100 documents (see the `"rows"` field) were returned to the classifier for further processing. Each document is presented with its fields as well as its relevancy score. This figure shows the first two relevant documents representing the terms *enterovirus- infectie* and *enterovirus viraal; infectie*. These terms were stored verbatim (as is).

More samples from the result set are shown in Table 5.5, where terms of any number of words match with the query at the beginning, in the middle, or at the end of their strings. These terms were added into the UMLS from various vocabulary sources and have at least two semantic types.

Our informal experiments showed that using all of the returned results leads to a low precision since the matching terms can be of any semantic type. To get a higher precision, we filter the terms to get a subset of terms most similar to the query term. For this purpose, we use the following filters:

number of words: For multi-word terms, filter out a term if the number of its words is not the same as the number of words in the query term.

head’s position: For Dutch term, filter out a term if the position of its head word is different from the position of the head word (if any) of the query term.

frequency: If there are more than three semantic types, take only the three most frequent ones.

[6] JSON (JavaScript Object Notation) is “a lightweight data-interchange format. It is easy for humans to read and write. It is easy for machines to parse and generate” (http://www.json.org/).

Following the heuristic for that query term, we are left with a subset of terms that consist of two words, e.g. *enterovirus echo, enterovirus; meningitis, enterovirus diseases, genus enterovirus*, etc. After counting the frequency of each semantic type in this subset, we get three most frequent ones, namely A1.1.3:*Virus* (28), B2.2.1.2.1:*Disease or Syndrome* (12), and A1.4.1.2.1.7:*Amino Acid, Peptide, or Protein* (3); numbers in the brackets are their frequencies. These semantic types are then returned by the classifier to label the query term *infectious enterovirus*. We decided not to take only the most frequent label, because in our experiments sometimes the best labels are in the second or third rank. By selecting a maximum of three labels, we expect to capture more correct labels for the next processing stage, i.e. relation extraction.

5.4.3 Temporal Filters

Temporal concepts that are important entities in medical relations are also found in the annotated corpus, for example, *tijdens de zwangerschap* 'during the pregnancy', *'s nachts* 'at night', and *per dag* 'daily'. There are 48 temporal terms among the 2000 terms being labeled. From this number, our classifier is only able to find 11 temporal terms from its database, such as, *'s morgens* 'at morning,' *'s avonds* 'at evening,' and *'s ochtends* 'at morning'. It failed to recognize easily predicted temporal terms such as *tijdens de zwangerschap* 'during the pregnancy', *na de operatie* 'after the operation', and *in korte tijd* 'in short time'. For these kinds of temporal terms, our classifier returns empty results.

To solve this problem, we investigate the unlabeled terms and then manually create a set of temporal filters that should be executed over terms that did not receive any label from the heuristic. These filters check if the unlabeled terms match with one of the regular expressions as shown in Figure 5.6. If that is the case, a *Temporal Concept* type of the UMLS with the semantic tree number *A2.1.1* will be assigned to the matched term.

5.5 Experiment and Results

We extract all known terms in the corpus (see Section 5.3) that have been annotated with concept types, which results in a set of 24,423 unique terms. For each of the terms, we apply the algorithm in Section 5.4.2 to obtain its matching UMLS labels.

Table 5.6 and 5.7 show 20 examples of the labeling results, each is for a set of single-word terms and a set of multi-word terms, respectively. The number signs (#) in the tables indicate the ranks of the terms according to their frequency in the corpus. The 'Corpus label' and 'UMLS label' columns present the classification results by the manual annotation and by our classifier, respectively. The 'Step' columns indicate at which step our classifier finds the label for the term. The values in the 'Eval' columns are determined automatically from the mapping. For example, the term *bloed* 'blood' is labeled with *Tissue* (A1.2.3.2) by our method, and since this semantic type has been mapped to *body_part*, the evaluation result will be 1 (correct).

These tables show that most of the terms in the examples are correctly labeled by our classifier. Moreover, compared to the manual labeling, our clas-

```
/^enkele /
/^voor /
/^na /
/^tijdens /
/dagen$/
/^bij de/
/^langdurig /
/^in de loop van /
/ jaar$/
/ jaren$/
/ tijd$/
/ minuten$/
/ uur$/
/ weken$/
/ uren$/
/ lang$/
/^vanaf /
/ halfuur$/
/ jarenlang$/
/ duur$/
/^in.*tijd$/
```

Figure 5.6: A set of temporal filters to classify temporal terms.

sifier returns more detailed labels. It is because our classifier uses all of the 135 UMLS Semantic Types as the target labels, while the manual labeling only uses 11 target labels. For example, terms annotated as *person* in the corpus are assigned with more detailed labels by our classifier, such as *kind* 'child' with *Family Group*, *kinderen* 'children' with *Age Group*, *patiënt* 'patient' with *Patient or Disabled Group*, and *vrouwen* 'woman' with *Population Group*.

However, an ambiguity occurs when human annotators have to decide among these two labels: *disease* or *disease_symptom*. For some terms, it is not clear to which labels they belong. For example, the term *angina pectoris* was labeled as a *disease* in the corpus. By definition, the term *angina pectoris*[7] is a *chest pain*, and *chest pain*[8] is a *symptom*. Therefore, the correct label for this term should be *disease_symptom*. This shows that these two labels are often interchangeable. Although our classifier has correctly labeled that term with *Sign or Symptom* (at step 3), for the evaluation purpose, we follow the manual labeling as the correct one.

[7] Angina pectoris, http://en.wikipedia.org/wiki/Angina_pectoris

[8] Chest pain, http://en.wikipedia.org/wiki/Chest_pain

#	Term	Manually assigned label	Automatically assigned label	Step	Eval
1	bloed ‘blood’	body_part	Tissue	1	1
2	huid ‘skin’	body_part	Body System	1	1
3	kind ‘child’	person	Family Group	3	1
4	kinderen ‘children’	person	Age Group	3	1
5	patiënt ‘patient’	person	Patient or Disabled Group	3	1
6	pijn ‘pain’	disease_symptom	Sign or Symptom	1	1
7	lichaam ‘body’	body_part	Human	3	0
8	vrouwen ‘women’	person	Population Group	1	1
9	patiënten ‘patients’	person	Patient or Disabled Group	3	1
10	hersenen ‘brains’	body_part	Body Part, Organ, or Organ Component	1	1
11	infectie ‘infection’	disease	Disease or Syndrome	1	1
12	hart ‘heart’	body_part	Body Part, Organ, or Organ Component	1	1
13	urine ‘urine’	body_part	Body Substance	3	1
14	longen ‘lungs’	body_part	Body Part, Organ, or Organ Component	1	1
15	spieren ‘muscles’	body_part	Tissue	1	1
16	arts ‘doctor’	person	Therapeutic or Preventive Procedure	1	0
17	rsi ‘rsi’	disease	Disease or Syndrome	3	1
18	bacterië ‘bacteria’	microorganism	Bacterium	1	1
19	ontsteking ‘inflammation’	disease	Pathologic Function	1	1
20	ernstige ‘serious’	disease_feature	Idea or Concept	3	0

Table 5.6: Examples of single-word terms classified manually using corpus labels and automatically by our method using the UMLS semantic types. The terms are ranked based on their frequency in the corpus.

#	Term	Manually assigned label	Automatically assigned label	Step	Eval
58	rode bloedcellen ‘red blood cells’	body_part	Cell	4	1
62	witte bloedcellen ‘white blood cells’	body_part	Cell	4	1
64	dunne darm ‘small intestine’	body_part	Body Part, Organ, or Organ Component	4	1
89	dikke darm ‘large intestine’	body_part	Body Part, Organ, or Organ Component	4	1
169	centraal zenuwstelsel ‘central nervous system’	body_part	Body System	4	1
188	tijdens de zwangerschap ‘during pregnancy’	duration	Temporal Concept	3	1
224	microscopisch onderzoek ‘microscopic examination’	method_of_diagnosis	Laboratory Procedure	3	1
229	angina pectoris ‘angina pectoris’	disease*	Sign or Symptom	3	0
235	diabetes mellitus ‘diabetes mellitus’	disease	Disease or Syndrome	3	1
250	jonge kinderen ‘young children’	person	Family Group	4	1
289	lichamelijk onderzoek ‘physical examination’	method_of_diagnosis	Health Care Activity	4	1
293	hart- en vaatziekten ‘cardiovascular disease’	disease	Disease or Syndrome	2	1
324	reumatoïde artritis ‘rheumatoid arthritis’	disease	Disease or Syndrome	3	1
339	’s nachts ‘at night’	duration	Temporal Concept	3	1
366	neutrofiele granulocyten ‘neutrophilic granulocytes’	body_part	Cell	3	1
396	twaalfvingerige darm ‘doudenum’	body_part	Body Part, Organ, or Organ Component	4	1
406	hoge koorts ‘high fever’	disease_symptom	Disease or Syndrome	4	0
425	na de geboorte ‘after birth’	duration	Temporal Concept	3	1
427	hoge bloeddruk ‘high blood pressure’	disease*	Disease or Syndrome	4	1
448	nefrotisch syndroom ‘nephrotic syndrome’	disease	Disease or Syndrome	4	1

Table 5.7: Examples of multi-word terms classified manually using corpus labels and automatically by our method using the UMLS semantic types. The terms are ranked based on their frequency in the corpus. The * sign indicates an incorrect labeling by human annotators.

	Frequency			
Step	**Single**	**Multi**	**All**	**%**
1	662	14	676	33.8
2	310	1	311	15.6
3	590	69	659	32.9
4	91	115	206	10.3
5	109	39	148	7.4
Total	1762	238	2000	100

Table 5.8: The numbers of the single-word and multi-word terms labeled at each step.

Such an ambiguity also occurs during labeling the term *hoge koorts* 'high fever'. By definition the term *koorts* 'fever' is a *medical symptom.* The manual labeling has correctly assigned the term the *disease_symptom* label. However, since all samples consisting of two words and having *koorts* as their head word, such as the sample terms *omsk-hemorragische koorts*, *faryngoconjunctivale koorts*, and *hemorragische koorts*, are labeled as *Disease or Syndrome*, our classifier will assign the term *hoge koorts* with this label (*Disease or Syndrome*) instead of with *Sign or Symptom.*

Most of the single-word terms in Table 5.6 are successfuly labeled by our classifier at step 1, which matches at the root form level of Dutch (or found in the Dutch part of the UMLS), and step 3, that matches at the root form level of English (or found in the English part of the UMLS). On the other hand, most of the multi-word terms in Table 5.7 are labeled at step 3 and step 4, that matches at the head word level of Dutch (or not found in the Dutch part of the UMLS).

To understand these findings, consider Table 5.8 that presents the numbers of single-word and multi-word terms labeled at each step. This table shows that most of the single-word terms in the corpus are known terms that can be found either in the Dutch part (662 terms at step 1 and 310 terms at step 3) or in the English part (590 terms at step 3) of the UMLS. On the other hand, the multi-word terms in the corpus are new terms for Dutch, where 69 terms (Step 3) and 39 terms (Step 5) are from the multilingual part, and 115 terms (Step 4) are from the same-head-word terms in Dutch. This shows that most of the new single-word terms in Dutch can be found in the English part through translation, while most of the new multi-word terms in Dutch are apparently also new terms for the English part. Table 5.8 also shows that in general, the use of translation has increased the number of unseen terms that can be labeled. Most of them have exact translations in English (32.9%) and some have similar samples in both languages (7.4%).

5.6 Evaluation

We take from candidate terms in the corpus a subset of the 2000 most frequent terms, which consists of 1762 single-word terms and 238 multi-word terms. We automatically evaluate their lables, which are generated by our labeling system, using term labeling information that have been extracted from the manually annotated corpus (see Section 5.3.4).

Since our classifier uses all of the 135 UMLS labels, whereas the IMIX corpus

	Precision		
Step	**Single %**	**Multi %**	**All %**
1	84.3	57.1	83.7
2	88.7	100.0	88.7
3	63.2	84.1	65.4
4	71.4	84.3	78.6
5	7.3	56.4	20.3
Avg	72.6	78.2	73.3

Table 5.9: Precision of the labeling on the single-word and multi-word terms at each step.

uses only 11 corpus-specific labels, we defined a mapping from the UMLS labels to the IMIX corpus labels (see Table 5.3). Note that each UMLS label was mapped to at most one corpus label. Evaluation is done on the basis of the highest ranked UMLS label.

The precision of our term labeling method is shown by Table 5.9. For the single-word terms, the precision of exact matching at root form and at string levels are relatively high (84.3% and 88.7%, respectively) compared to the precision of the exact translation (63.2%). This shows that our classifier performs better in finding similar samples from the Dutch part than from the English part for the single-word terms.

For new single-word terms that do not have any translation, labels are selected from multi-word samples that match at the head word level (step 4). At this step, our classifier demonstrates a 71.4% precision. The absence of syntactic information (*head*) during the sample selection causes the precision to drip signicantly (7.3%). This shows that for new single-word terms, head words are very important in finding similar samples.

As for the multi-word terms, it is surprising that the precision at the root form level is significantly lower (57.1%) than to the precision at the same level for the single-word terms (84.3%). Intuitively, they should demonstrate a similar performance. When we looked at the labeling results, we found that 5 out of 6 errors occurred on the terms '*vitamine* X', e.g. *vitamine K*, *vitamine B12*, and *vitamine D*. All of these terms are labeled as *body_part* in the corpus, while in the UMLS, they, as well as other laboratory substances (e.g. *albumine*, *prolactine*, and *oxytocine*), are labeled as *Laboratory Procedure* (B1.3.1.1). Initially, we did not map *Laboratory Procedure* to *body_part* because their meanings are unlikely to be similar, and moreover, this label has been mapped to *treatment*. One may question this mismatch: Which one is correct? We consider both labelings correct with respect to the set of semantic types they use. Classifying the term '*vitamine* X' into *body_part* is the best match we can get using the corpus labels, although it is not always true that '*vitamine* X' is a *body_part*. Therefore, labeling done in the UMLS is better, since with the 135 semantic types, it is possible to classify terms like '*vitamine* X' to more fine-grained labels. This finding suggests us that to get a better accuracy in classifying terms in the corpus, 11 semantic types are not enough.

Another error, 1 out of the 6 errors, occurs with the term *anorexia nervosa*. This terms was labeled as *disease* in the corpus, while in UMLS it is labeled *Sign or Symptom*. It is another ambiguity that we found, since this term will

Label	Total	Precision %
disease_symptom	138	80.4
disease	616	80.0
microorganism	42	76.2
body_part	618	74.3
duration	48	72.9
method_of_diagnosis	54	70.4
treatment	207	69.1
person	104	66.3
bodily_function	99	61.6
person_feature	28	35.7
disease_feature	46	34.8

Table 5.10: The numbers of terms based on labels and their precision.

intuitively be classified as a 'disease' by a human annotator, although technically it is a 'diagnosis' of a disease. Thus, if we adjust our mapping reflecting these 'mostly ambiguity' errors, the performance for the multi-word terms would be higher.

The precision of our method in labeling new multi-word terms is significantly higher than for labeling new single-word terms. It has a 84.1% precision at the exact translation level, a 84.3% precision at the head word level (step 4), and a 56.4% precision at the word level (step 5). The use of syntactic information (head) at step 4 gives better results compared to the absence of this information at step 5; this result is similar to that for the single-word terms. At step 5, the better performance on the multi-word terms side is contributed by the 'number of words' filter. This filter will only allow samples with the same number of words as the query term, which is hypothesized to result in samples similar to the query terms. For the single-word terms, we cannot apply this filter at step 5 since it will require exact matches at the string level (step 2), and as a consequence, spurious samples will pop up during the sample selection and the performance will decrease.

Table 5.10 shows the number of terms and their precision according to the assigned labels. Despite the ambiguity of the labeling between the corpus labels *disease_symptom* and *disease*, the precision at these labels are the highest (80.4% and 80%, respectively). The precision for other labels such as *microorganism*, *body_part*, and *duration* are also relatively high. The lowest precision is achieved by *person_feature* and *disease_feature*. If we look at the terms assigned with these labels, we get the impression that most of the 'feature' terms are adjectives, such as *dodelijk* 'deadly', *besmettelijk* 'contagious', and *ernstiger* 'serious'. These terms are labeled as *disease_feature* in the corpus. And since an adjective is usually not a *head*, the labeling of these terms has not benefitted from the 'head' filter. Again, this result shows the importance of *head words* in labeling terms.

It is difficult to compare the performance of our method with other methods, since the corpus and the labels we use are not shared with other works. The only work using the same corpus is Canisius et al. (2006), in which the authors use a machine learning approach to train a concept classifier. They do not use external resources, but instead try to learn the classification from (a subset of) the corpus itself which results in an accuracy of 68.9%.

5.7 Summary

We have described our approach to labeling medical terms using an existing terminology, i.e. the UMLS. This work is aimed at answering research question #5: to what extent labels in the UMLS can be used to classify unseen terms in a particular language, such as Dutch.

Before applying our method, we index entries of terms and their labels extracted from the UMLS using an IR tool aimed at getting a high performance with respect to the processing speed, and most importantly, applying linguistic analyses during the indexing and the searching phases. The linguistic analyses include stemming for English terms and matching head words for Dutch terms.

Our method uses an example-based strategy by collecting a set of sample terms which are similar to the query term. To get the similar terms, we apply a heuristic consisting of 5 steps, where in each step a query containing specific parameters is sent to the IR engine. If there is no similar term returned at a step, the heuristic will proceed with the next step, until a set of similar terms is returned or no more steps remain.

Sample terms returned at later steps, e.g. at step 4 and 5, are usually less similar to the query term compared to sample terms returned at the earlier steps. Especially for these less similar terms, we apply another heuristic aimed at filtering out terms that do not match with at least one of the following filters: number of words, head's position, and frequency.

We evaluated this method in a task of labeling the 2000 most frequent terms extracted from the IMIX medical corpus, that has been annotated with 11 concept types, using the 135 semantic types from UMLS. Our evaluation shows that most of the single-word terms in the corpus are *known* terms with respect to the UMLS terms. From these terms, 33.5% are new Dutch terms whose labels can be found in their exact translations. On the other hand, most of the multi-word terms in the corpus are *new* terms, of which 29.4% are labeled through exact translation. On average, our classifier results in a 73.3% precision.

At both types of query terms (single- and multi-words terms), *head words* play an important role in selecting sample terms that have a high similarity to the query terms. Based on the results in Step 4 of the algorithm, we can draw the following conclusion: terms which have the same head word as the query term tend to share the same label with the query term. This conclusion is also supported by our finding that the precision of labeling terms tagged with *adjective* is low, since these terms usually do not have any function as head words.

For the multi-word query terms, the number of words is also a good indicator for finding similar terms. Since we do not apply this filter to the single-word query terms, these terms are suffering from the non-similar samples of any number of words. These findings are supported by the results at step 5.

As a wrap up, we can draw a conclusion that, thanks to the translation, head words, surface length, and frequency, we can use a multilingual terminology containing terms and their labels to classify unseen terms of a particular language. Our experiment results also showed that a list of detailed classes, such as the UMLS Semantic Types, is necessary to assign more accurate classes to medical terms. Our classification method is aimed at achieving this goal by using the whole UMLS Semantic Types as target classes. The potential effect of this approach is that our method can be used to populate and label new terms and then add them to the existing terminology.