Studienarbeit

Computing Semantic Similarity Between Medical Learning Objectives Across Catalogues

Eingereicht von:

István Sárándi

Matrikelnummer: 328707

Betreuer: Dr. C. Spreckelsen

Eingereicht: 2. März 2016

Abstract. The standardization of medical learning objectives across Germany's medical universities requires that each university finds a correspondence between their learning objective catalogue and the new central, standard one. In practice, this boils down to comparing phrases and sentences across the catalogues by semantic similarity. In order to make this process faster than manual work, we propose an automated system to compute such similarities and to return a list of the most likely matches for an input query. The system relies on vector space representations of the learning objectives where the representation makes use of medical domain knowledge in the Medical Subject Headings (MeSH) database. In more detail, the learning objectives and the MeSH entries are first represented as weighted bag-of-words vectors. Then, for each learning objective, the cosine similarity is computed between it and every MeSH entry, producing a list of numbers, which is then interpreted as a new vector, thus a transformed representation of the learning objective in a so-called MeSH space. Finally, we can formalize the semantic similarity between two learning objectives as their cosine similarity in MeSH space.

1 Introduction

The study is concerned with the task of finding semantically corresponding medical learning objectives in two catalogues. Learning objectives are specific and measurable requirements (skills and pieces of knowledge) that medical students need to know for different university examinations. The need for solving this inter-catalogue matching problem arose due to the development of a new unified and standardized national medical LO catalogue in Germany. This catalogue is the National Competency-based Catalogue of Learning Objectives for Undergraduate Medical Education (NKLM) of the German Medical Faculty Association (MFT) and the German Medical Association[3].

Medical universities such as the medical faculty of RWTH Aachen University already have their own set of such learning objective descriptions in place[12]. These representations need to be linked to the NKLM for standardization purposes. Matching them by hand without any computer assistance would require a long time of monotonous work. This time (and the associated costs) can be reduced if the process is automated, at least partially.

The goal of this study is to create a conceptual framework for this automation and to propose a simple but effective statistical scheme for evaluating

semantic similarity. Furthermore, we will also review some of the more sophisticated approaches in recent literature that could be incorporated into a more advanced version of the system.

1.1 Semantic similarity

Evaluating semantic similarity between expressions or sentences is still an unsolved problem in computer science. As of today, there are no methods that could understand and reason about natural language on a human-equivalent level. A true solution to the problem will have to build upon several fields including text mining, artificial intelligence, machine learning and computational linguistics. Part of the challenge is to formalize what "meaning" actually is or ought to be in a computerized system.

A prominent and practically useful theory of meaning, called distributional semantics, claims that the meaning of a word or expression can be best captured by its usage patterns, i.e. in which contexts is it customary to use it. This can be approximated in a statistical way by computing word usage frequencies in large corpora of text. Such frequencies and co-occurrences will be the basis of the approach proposed in this study as well.

Anything that we want to manipulate in a computer must have a well-defined representation. Most language technology research represents concepts and meaning as vectors, in other words lists of real numbers. This advantageous since the mathematics of vectors and vector transformations are well understood and developed and the vector operations are simple to implement. Vector representation can also be the bridge to the field of machine learning, whose methods typically require the input data to consist of vectors.

1.2 Language processing

The input text in our case is written in German. To handle German text, we need several preprocessing steps due to the inflecting (fusional) nature of the language: different suffixes and vowel alterations are used to express the grammatical role of words. Therefore, word stemming will be necessary in

order to concentrate on the meaning of words rather than their place in the sentence. Additionally, German orthography rules give rise to compound words that need to be split to be able to better analyze their meaning.

Recognizing the grammatical structure of sentences (parsing) can be useful when looking for the most important words, although in our case the medical LO descriptions are generally simple sentences where a bag-of-words representation retains most of the semantics. Therefore, the parsing component in the system will be optional.

To successfully solve any automation task, we have to carefully analyze the domain in which we work, since a general solution that would also be effective outside the scope of the requirements will often take much more effort than actually needed for the task at hand. In our case, we know that the domain in which we have to work is medicine. Several databases of medical concepts and expressions exist (dictionaries, thesauri, taxonomies, ontologies) and by using them we can import medical domain knowledge into the system. One such example is called Medical Subject Headings (MeSH), created and maintained by the United States National Library of Medicine. It is a controlled vocabulary, organized into a directed acyclic graph (DAG) structure and includes synonyms and alternative formulations for most concepts. Given that a German language edition of MeSH is also available, it will be a useful component for our system. General ontologies and lexical databases, e.g. WordNet, usually do not include enough medical terms to be useful for our application.

1.3 A brief description of the proposal

The main idea of the proposed system is to represent the meaning of LOs by vectors in a high-dimensional space that we shall call "MeSH space". Each dimension in this space corresponds to a MeSH heading and the value along the dimension describes the relevance of the heading for that particular LO. We can use standard distance metrics or similarity measures to find the best corresponding LOs from the other catalogue. By returning a list of multiple neighboring LOs we can let a human expert decide about the final correspon-

dence assignment. This way the system will not be fully automatic, but the amount of work required will be still significantly reduced, since the human expert does not have to search over the whole database to find a matching LO. There is a tradeoff to the list of returned LOs. It should be long enough to include the real solution but short enough to be comfortable to read through.

The study is structured as follows. In Section 2, we present some related work in computational semantics. Then in Section 3, we describe the data sources that we used for testing the proposed method. The methods themselves are presented in Section 4. Section 5 contains a few example results for qualitative analysis (detailed quantitative analysis will become possible when a benchmark dataset is created with the assistance of experts). Finally, Section 6 gives a summary and discusses further possibilities for adjusting or improving the system.

2 Related work

The fields of information retrieval, natural language processing and text mining have developed many useful tools for handling and processing free-text, natural language data. Semantic methods have been applied for machine translation, document retrieval and natural language interfaces.

Most approaches work based on the so-called distributional theory of meaning[11], which claims that the distribution of a word or phrase across a large corpus is the key to capturing its meaning. If two words appear in similar contexts, their meaning is judged to be similar by this approach. Most of the early results in statistical natural language processing was based on the setup of information retrieval. This meant that the data was a set of larger text documents, such as news articles, and the systems had to perform query-based search. This naturally lead to the analysis of word-document matrices, which contain the information about which words appear in each document. One such approach is Latent Semantic Indexing[5].

A more explicit representation of meaning is used in ontologies and digital thesauri, which encode hierarchical and other relationships between words in a

systematic fashion. Typically, creating and maintaining such databases requires a large amount of manual work, therefore there are only a few large-scale projects doing such work, the best known of which is WordNet[7], a general, English language word database. In the medical field, the Medical Subject Headings (MeSH) is a well-known controlled vocabulary, that is available in several languages, including German[6].

Several approaches for semantic similarity computation rely on online services. One such approach is the Google similarity distance[1], which compares the number of search results returned for two strings separately and used in conjunction.

Besides the classic probabilistic methods based on computing statistics on the document-word matrix, there has been an increased interest recently in neural networks to produce vector representations for words. The usual example to illustrate the idea behind these systems, is the $king + (woman - man) \approx queen$ equation.[10].

3 Materials

In this section we present the source of the datasets and word lists used in this study.

3.1 NKLM

The National Competency-Based Learning Objectives for Undergraduate Medical Education (German: Nationaler Kompetenzbasierter Lernzielkatalog Medizin, NKLM) of the German Medical Faculty Association (MFT) and the German Medical Association will be a catalogue of medical learning objectives, aimed at standardizing the curricula of medical faculties throughout Germany. See [3] and [13] for more information on this catalogue from the medical education system point of view. For our purposes, this database is a set of records organized into a hierarchy of chapters and subchapters (4 levels). Each record consists of a main description part (required), an examples part (optional),

references to other chapters (optional), an associated disease (optional). The texts are free-form German language descriptions. For this study, we are using a preliminary version of the NKLM dataset, consisting of 2213 records. An example record contains the following components:

1. **Chapter:** 12.2.4.4
2. **Description:** die Regulation von Enzymen durch allosterische Regulatoren, posttranslationale Modifikationen und limitierte Proteolyse erklären.
3. **Examples:** Stoffwechselregulation; Wirkung von Insulin; Komplementaktivierung
4. **Cross-references:** Blutgerinnungskaskase; Fibrinolyse; 16; Pharmakotherapie
5. **Associated diseases:** Pertussis; Cholera; Diphtherie

3.2 Aachen Catalogue of Learning Objectives

The Aachen Catalogue of Learning Objectives (ACLO) was created at the Uniklinik RWTH Aachen over the course of 25 months. It is managed in a custom developed social semantic web platform built for collaborative curriculum mapping.[12]

Let us look at an example record of the ACLO dataset:

1. Diagnostik - Abdomenuntersuchung - 15
2. Gastroenterologie und Stoffwechselkrankheiten
3. Abdomenuntersuchung
4. Der Studierende soll
5. aus den Befunden einer gezielten Abdomenuntersuchung die Differentialdiagnosen der Befunde
6. erklären
7. bewerten

The first three components describe the context of the learning objective including the medical field of study. The rest of the components describe the learning objective itself in a sentence with a rigid grammatical structure. The

4th component is typically "Der Studierende soll" (The student should) the 6th (and possibly 7th) part is a verb and the 5th part contains all other parts of the sentence.

In our study, we will use the concatenation of the 3rd and 5th components for matching. The 5th component is obviously necessary as it is the longest part, which usually contains most of the important words. The 1st and 2nd components are rather broad categories and are implied by the words in the 3rd and 5th components. The 3rd component is also required as there are records where the 5th component is a rather general description that can not be understood without context. An example for such 5th component is "die entsprechenden differentialdiagnostischen Untersuchungen und Symptome" (the corresponding differential diagnostic examinations and symptoms). Without the 3rd component "Asthma-Anfall (Notfallmeidzin)" (Asthma attack, emergency medicine), the record would not be understandable.

3.3 Medical Subject Headings

Medical Subject Headings (MeSH) is a medical thesaurus or controlled vocabulary, created by the United States National Library of Medicine in 1960. Its main purpose is to allow categorization of the medical literature, to make search and information retrieval more efficient.[6]

We can also make use of this vast knowledge base for our purposes in this study. The headings represent medical concepts, organs and diseases that are central to interpreting the meaning of a learning objective. By making use of its descriptions and structure, we can use medical expert knowledge in our system instead of only relying on data.

3.4 Additional German language resources

Our method requires collecting a large dataset of German language words for handling compound and derived words. In addition to the words contained in NKLM, ACLO and MeSH, we will use the following sources.

The first source is the GermaNet 8.0 compound splitting database created at the University of Tübingen.[4]

The second source is Wiktionary, a free online dictionary hosted by the Wikimedia Foundation and edited by volunteers. The dump of the Wiktionary database can be downloaded. It is available in several formats (including XML), but the plain-text version is simplest to process. It contains one line per dictionary entry and the German words that are relevant to us can be extracted by a simple regular expression. We match any German adjective, noun, verb or adverb that is not simply a derived form.

Further, we include the database of jWordSplitter, a Java-based German compound word splitter library created by Daniel Naber.[9] This is already provided as a list of word stems. We also include the words from Morphy Mapping, a key-value database of derived words and the corresponding stem also created by Daniel Naber.[8]

This amounts to a set of approximately 460000 words.

4 Methods

As mentioned in the introduction, semantic similarity of LOs will be calculated in a vector space called *MeSH space*, where the dimensions correspond to MeSH headings and the vectors' coordinates along each dimension heuristically approximate how much the heading corresponding to the dimension is semantically relevant to the LO. We will now give an overview of how this MeSH vector can be created.

The idea is to first introduce a very high-dimensional *word space*, where each dimension corresponds to a possible German language word (more closely defined later). Each MeSH heading and each LO will be represented as a vector in this word space first. The vectors' coordinates will approximate the relevance of the given word to the heading or to the LO. Thus, we do not only consider whether a word is part of the LO (or the description of the heading) but will also consider how "strongly" it is part of it, giving us a finer-grained representation than the usual bag-of-words model.

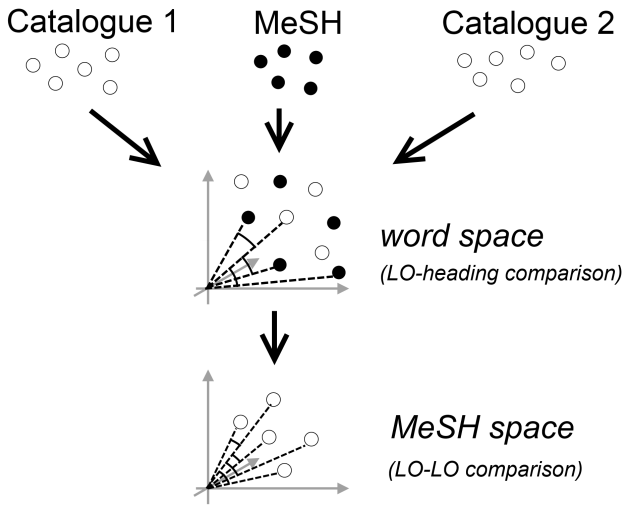

Fig. 1. The overview of the method. LOs and MeSH headings are first represented in word space. Then based on LO-heading similarities, the LOs are represented in MeSH space, where LO-LO comparisons can be performed.

The MeSH space representation of an LO will then be constructed by calculating a similarity score between the LO and each MeSH heading in word space.

We have to expand on what "words" should exactly mean, how they can be extracted from LOs (or heading descriptions) and how their relevance can be defined and calculated.

A word (a dimension in word space) should intuitively be a "meaningful fragment" of the multiword expression, meaningful to for our medical semantic purposes. Inflection and certain suffixes are simply used for embedding the word in a sentence and are therefore not relevant to the core meaning. Thus, a word should be considered as a prototypical, abstract entity representing all its various inflected forms. On the other hand, our set of such abstract words should include compound words as well, since they can carry additional meaning above the fact that their constituent parts are present. However, each time a compound is present, its constituents' presence should also be recognized. For example, if an LO includes "Kalziumstoffwechselstörungen", the extracted fragments should be:

- Kalzium
- Kalziumstoff
- Kalziumstoffwechsel
- Kalziumstoffwechselstörung
- Stoff
- Stoffwechsel
- Stoffwechselstörung
- Wechsel
- Wechselstörung
- Störung

Sometimes we get unwanted combinations as a side effect, such as "Kalziumstoff" or "Wechselstörung", but this effect does not lead to serious problems, since these pseudo-words will not come up often so similarities will not depend strongly on them.

In the next section, we will see the (somewhat tedious) details of how these fragments can be extracted. The relevance score of such a fragment to the whole LO will be calculated to approximate the "amount of meaning" or relevance that it carries in the context of the other parts of the LO. Importance will be measured based on the well-known and widely used inverse document frequency weight (idf) (more details later).

4.1 Extracting fragments

Purpose: In this step we take a phrase or sentence string and extract from it the units that carry meaning, i.e. individual words and compound components without regard to inflection.

To extract such fragments from a phrase, we need to be able to tell where compound boundaries are, and we need to be able to discard inflection or suffixes such as the plural that are meaningless for our current purposes.

Compound splitting requires us to collect a set of words (*splitter words*) that are expected as subunits of a compound word. We will use the library jWordSplitter, which has its own database. This is, however, a general-purpose

database that lacks words needed in our special medical context. Therefore the set of words has to be expanded. The Materials section already described the source of these new words. We now discuss how we preprocess this set of words to get a set of useful splitter words for use in jWordSplitter.

Preprocessing of the set of splitter words for compound splitting *Purpose:* In this step we take a set of words and filter out the compounds. Words are regarded as compounds if they can be composed of other words that are included in this set.

First some simple cleaning steps are undertaken

- Convert to lowercase (the final system will be able to distinguish letter case, since it is important in German, but the compound splitter does not need such a distinction)
- Trim punctuation marks
- Remove words under length of 3 characters
- Remove words containing no vowels
- Remove words over the length of 5 that end with -es or -s for which the word set contains a version without this ending. This is needed for proper recognition of compound words, as these -s- pieces can be "compound glue" in German words.

We also synthesize some additional suffixed words based on specific rules, such as for each word ending in -lich, we add the corresponding word ending in -lichkeit.

It is crucial to remove compound words from the database of the compound splitter itself. If, for example, "Blutdruck" were included among the splitter words, jWordSplitter would not fully split our words to its basic constituents, as it would accept "Blutdruckmessung" as simply ("Blutdruck", "Messung") without revealing the further inner structure.

This compound removal from the splitter words is done with jWordSplitter itself. The set of splitter words is first initialized to still include compounds. Then we test each word for compoundness one-by-one. The trick is that before

testing the compoundness of a word, we temporarily remove it from the splitter set. If we get more than 1 component, then the word is compound and it is discarded from the splitter set.

After doing this, we now have a good set of splitter words to be used with jWordSplitter. The next task for fragment extraction is the creation of abstract, inflection-agnostic word stems.

Handling derived words *Purpose:* The German language uses inflection and suffixes to express grammatical functions of words. This can be considered noise for our purposes in making sense of the overall meaning of an LO and this step is used to make the fragment extraction process independent on how the words are inflected or suffixed in the particular sentence.

The straightforward way to proceed would be to use a stemmer algorithm such as Snowball Stemmer to remove endings. Unfortunately Snowball Stemmer is too aggressive in removing endings and it is therefore unusable in our case. The solution will be a combination of two methods.

The first method is a pre-existing fixed list of (inflected, uninflected) pairs called Morphy Mapping, compiled by Daniel Naber. This mapping does not strip all endings, only one. Therefore we apply a "transitive iteration" on this mapping: if the pairs $a \mapsto b$ and $b \mapsto c$ exist in the mapping, we modify it to $a \mapsto c$. Additionally, we do another such iteration for words that are capitalized, and are not included in the words of Wiktionary. This is aimed at the cases such as: $Einfacheren \mapsto Einfachere$, $einfachere \mapsto einfach$.

However, even at 368165 pairs, this Morphy Mapping list is not long enough to include all words that we encounter in the medical context. Therefore, a second method will be needed. It is hard to decide whether a new, unknown word is inflected or just happens to end with such a sequence of characters by coincidence. To solve this, we will not simply trim endings off a word, but create *derived word sets*. A derived word set includes the inflected forms (endings -er, -e, -es, -en, -em, -n, -s) of a given word. Formulated otherwise: we represent each word stem not as a concrete uninflected string but as a *set* of strings, containing its possible inflected variants. A map data structure is

then used to convert a particular string (inflected word form) to the abstract stem object (standing for a set of forms).

In practice we manage a single central mapping $\Phi : \Sigma^* \to 2^{\Sigma^*}$. This mapping is expanded each time we encounter a new word whose stem we do not yet know: we generate versions with added inflectional endings and, additionally, if the word seems to end in an inflectional ending, we also generate the supposed remaining part after splitting the seemingly inflectional ending.

Now we have the tools to split compounds and to get the abstract stem of a word. We should now proceed to how a multiword expression (LO or MeSH entry) can be processed to yield a list of scored fragments.

Analysis of multiword expressions (LOs, MeSH entries) *Purpose:* In this part we describe how the fragments are extracted in practice, using the compound splitter words and the stemming approach that we outlined above.

We first split the multiword expression to words at whitespace or punctuation. We discard stopwords (grammatical words that have no meaning for us). We then split each word to its constituent parts (in case it is compound) by the following steps:

- Use Morphy Mapping to remove eventual inflection or suffixes. Decide on capitalization. If the word was lowercase originally and that lowercase word is among the keys in Morphy Mapping, then it is kept lowercase. Else, if the capitalized version is in Morphy Mapping then we make it capitalized. Otherwise we keep the word as it was.
- Use jWordSplitter to split the word to its constituent parts
- Split the resulting words further at hyphens
- Create two lists of the constituent parts. One that contains all inflections (and also "compound glues" such as -s-); and one that contains the Morphy Mapped (uninflected) versions of the words. This is done the same way as the first step above describes. Capitalization is also decided according to the above rules, the only difference is that now the "original" capitalization is taken from the whole word's beginning.

In the case of "Gesundheitsdienste" the first resulting list is ("Gesundheits", "Dienste"), the second is ("Gesundheit", "Dienst"). We now create all the fragments, as we have seen on the example of "Kalziumstoffwechselstörungen" before. All possible compounds are extracted. In case of each compound, we take the form from the first list (with "glue") for all pieces except for the last, where we take the form from the second list (uninflected). This results in:

- Gesundheit
- Gesundheitsdienst
- Dienst

All these fragments are then put into the central Φ mapping described above to handle different inflectional forms. For example, if there is not yet a key in this mapping for "Gesundheit", we create an abstract word stem object for it, and map to this abstract object from a set of synthesized inflected forms. We do not need to give much grammatical care here, it is enough to use brute force and simply create many possibilities, such as:

- Gesundheit
- Gesundheite
- Gesundheits
- Gesundheiten
- Gesundheitn
- Gesundheiter
- Gesundheitem

Even though we also create nonsense words this way, the important point is that the right ones are generated as well, and the false ones do not have any negative impact on performance. Now all these created forms point in Φ to the abstract word object representing this set of inflected forms.

At this stage, we have a list of fragments in their abstract stemmed form. The next step is to give relevance weights to them according to "how much

of the meaning they carry" in the overall multiword expression. These weights will be the coordinates of the vectors that represent the expression in word space.

4.2 Weighting the fragments in each expression

Purpose: Having extracted the word fragments from an sentence or phrase, we could use this directly for a binary bag-of-words vector representation. However, this would disregard the importance of each fragment in the sentence. Therefore, we introduce a weighting scheme that lets the system focus more on the relevant, meaning-carrying word fragments.

These relevance weights are based on the inverse document frequency (idf) factor, which has been successfully applied throughout text mining and information retrieval. In general, the idf weight of a word w is calculated based on its frequency in a set of documents D:

$$freq(w, D) = \frac{\left|\left\{w \in d \;\middle|\; d \in D\right\}\right|}{|D|}$$
$$idf(w, D) = \log \frac{1}{freq(w, D)}$$

Intuitively, the idf weight expresses how rare a word is, with the implicit assumption that rarely occurring words are more important than frequently occurring ones.

The frequencies are computed over the union of ACLO and NKLM for the LOs and according to MeSH frequencies for MeSH entries.

The relevance weight $relevance(w, d)$ of a fragment w in an expression d is the product of its general importance (idf weight) and its relative importance inside the expression. The relative importance $relativeImp(w, d)$ is determined using the already computed idf factors.

We would like to determine how much importance (modeled by idf weights) w has compared to the total importance in d. The total importance can be defined as the importance of w plus the sum of the importances of maximal

fragments outside w. For example, if we consider $w =$ "$Stoffwechsel$" and $d =$ "$Kalziumstoffwechselstoerungen$", the maximal fragments outside w would be "Kalzium" and "Störung". If $w =$ "$Kalzium$", the only maximal fragment outside w is "Stoffwechselstörung".

$$relevance(w, d, D) = \frac{idf^2(w, D)}{idf(w, D) + \sum_{\substack{v \in d \setminus w, \\ v\, maximal}} idf(v, D)}$$

We have to do an additional step in case of MeSH headings. Since they are not single sentences but have multiple entries (synonymous formulations), we have a different word-space vector for each entry. However, we would like to have a single word-space representation for a MeSH heading. A straightforward way to create that single representation would be to take the average of the entry vectors. But each entry of the heading is a full description on its own right, meaning that any fragment it contains is at least as relevant to the heading as it is to that particular entry formulation. This consideration leads to a coordinate-wise maximum operation over the vectors of entries in fragment space. Geometrically, this corresponds to taking the axis-aligned bounding box (bounding hypercube) of the entries' points in word space and choosing the vertex which is furthest from the origin.

4.3 Comparing learning objectives in MeSH space

Purpose: Having arrived at a representation of each LO and MeSH heading in word space, we will now transform the LOs from this word space to a new space where they will be compared for similarity.

This step builds upon ideas from both the support vector machine in machine learning and semantic mapping[2], a dimensionality reduction procedure used in text mining.

The support vector machine is a classifier and at test-time it works by projecting the input vector onto a small set of support vectors (or calculating

a similarity with them in case of the kernelized version), then taking a weighted sum of the resulting values and the classification result is based on thresholding this weighted sum.

In semantic mapping, words are clustered to topics in a *document space* (where each dimension stands for a document and words are represented in the space according to how many times they occur in each document). Then the topics are represented in *word space* (each topic is represented in the space according to how much a given word belongs to the topic), thus the topics take a similar role as support vectors in SVMs. Then each document is projected onto the topic vectors in word space (a document in word space represents how many times the words occur in it). Finally, the dimensionality reduced vectors are obtained by representing the documents in *topic space*, where a document is represented according to the projection values that were obtained.

Our idea will be quite similar but with one big difference: the analogues of topics or support vectors will not be chosen by clustering or learning, they will be the MeSH headings represented in word space. The rationale is that we have an a priori assumption that MeSH headings represent topics that are relevant in our medical setting. We also compute a similarity score instead of a simple projection (just like in a kernel method) between the LO and the heading vector. For the similarity measure, we choose the cosine similarity due to its simplicity and its successful applications throughout text mining.

$$cosSim(v_1, v_2) = \cos(\angle(v_1, v_2)) = \frac{\langle v_1, v_2 \rangle}{\|v_1\| \, \|v_2\|}$$

At this stage we now have a representation of each LO as a vector in MeSH space. The final similarity measure between LOs will be (again) the cosine similarity between these vectors.

To get the best matches in LO dataset D_1 for a query q in LO dataset D_2, we do a k-nearest neighbor search: we first compute the cosine similarities between q and every LO in D_2 and choose the top k most similar ones.

5 Results

Unfortunately the datasets are not annotated and therefore quantitative evaluation is not possible, however we can present some example matching scores as calculated with the Java implementation of the above described method. We will use the learning objectives from ACLO as queries and find matching LOs in the NKLM catalogue. Not all queries return high quality matches, therefore let us inspect some results where the similarity is above 0.8.

- Query: Aufbau und Funktionen der Extremitäten erklären.
  - Spezielle Unfallchirurgie: Untere Extremität — die klinischen Untersuchungen und Funktionstestungen der unteren Extremität erklären (0.83)
  - Spezielle Unfallchirurgie: Obere Extremität — die klinischen Untersuchungen und Funktionstestungen der oberen Extremität erklären (0.82)

- Query: Bewegungsstörungen und ungewollte Bewegungen
  - Bewegungsstörungen — den Begriff Bewegungsstörung und die Hauptformen von Bewegungsstörungen erklären (0.80)
  - Neuropathologie der Bewegungsstörungen — die exemplarischen Gendefekte, die zu SCA führen erklären (0.80)
  - Neuropathologie der Bewegungsstörungen — die charakteristischen histopathologischen Veränderungen bei SCA erklären (0.80)

- Query: die Maßnahmen für den Patiententransport erläutern.
  - Patiententransport (Notfallmedizin) — die Maßnahmen für den Patiententransport (Überprüfung Transportfähigkeit, erforderliches Monitoring, Unterschied innerklinisch/außerklinisch etc.) erklären (0.8641346797769653)
  - Epithelialer Transport — Lokalisation und Triebkräfte des Transportes von Wasser und NaCl erklären (0.844464780405614)

- Query: die an der Regulation des arteriellen Blutdrucks beteiligten Prinzipien und Mechanismen erklären.

- Hämodynamik — den mittleren arteriellen Blutdruck an Hand des systolischen und diastolischen Blutdrucks erklären (0.99)
- "Allgemeines, Nicht-medikamentöse Therapie" — die (Risiko-adaptierten) Blutdruck Zielwerte bei verschiedenen Begleiterkrankungen erklären (0.92)
- Physiologische Grundlagen des Kreislaufs — ausgewählte Prinzipien langfristiger Blutdruckregulation erklären (0.92)
- Physiologische Grundlagen des Kreislaufs — die Bedeutung des arteriellen Blutdruckes bei kurz- und langfristigen Störungen erklären (0.92)
- Physiologische Grundlagen des Kreislaufs — wichtige Prinzipien für die kurzfristige, insbesonders die korrektive Regulation des arteriellen Blutdrucks, erklären (0.92)
- Antiarrhythmika und HRST (Pharmakologie) — die Risikofaktoren für Hypertonie mit Anpassung der Blutdruckgrenzwerte erklären (0.92)
- Makroskopie (Anatomie) — die Grundprinzipien der Regulation von Blutdruck und Organdurchblutung erklären (0.91)

6 Conclusion and further research

Qualitatively, the results appear useful when the system finds matches with a high similarity score, but there are LOs to which no such match can be found. This could be due to an actual lack of corresponding LO in the other catalogue or the correspondence is too subtle to be discovered by this simple statistical method.

In this study we have presented a method to measure semantic similarity between free-text German language medical learning objectives that are specific pieces of knowledge that are required to be learned by medical students. By representing them in a so-called MeSH space, we capture the domain-relevant, medical meaning of learning objectives. The method is based on a combination of tried and tested ideas in natural language processing, including vector space representations, idf weighting and the cosine similarity function. Such a system can be of practical use when trying to match learning objectives across

catalogues as it can reduce the number of pairing that one has to manually check.

This relatively simple system could be improved in several respects. We could introduce a more refined concept of word importance instead of simply using the idf weight, perhaps based on how specific the word is for the medical terminology. This could work because a rare word can still be non-medical and therefore contribute less to the medical meaning of a phrase. Another direction of potential improvement is more sophisticated use of the MeSH hierarchy graph. The neural networks that have been developed in the last few years are also worth investigating for this use case.

In order to make further research progress quantifiable, it would be beneficial to manually assemble a benchmark dataset consisting of as many matching pairs of learning objectives as practicality allows.

References

1. R. L. Cilibrasi and P. Vitanyi. The google similarity distance. *Knowledge and Data Engineering, IEEE Transactions on*, 19(3):370–383, 2007.
2. R. Corrêa and T. Ludermir. Dimensionality Reduction of very large document collections by Semantic Mapping. *6th Workshop on Self-Organizing Maps*, 2007.
3. E. G. Hahn and M. R. Fischer. National Competence-Based Learning Objectives for Undergraduate Medical Education (NKLM) in Germany: Cooperation of the Association for Medical Education (GMA) and the Association of Medical Faculties in Germany (MFT). *GMS Z Med Ausbild*, 26(3):2009–2026, 2009.
4. V. Henrich and E. Hinrichs. Determining Immediate Constituents of Compounds in GermaNet. In *Proceedings of Recent Advances in Natural Language Processing*, pages 420–426, Hissar, Bulgaria, 2011.
5. T. K. Landauer, P. W. Foltz, and D. Laham. An introduction to latent semantic analysis. *Discourse processes*, 25(2-3):259–284, 1998.
6. C. E. Lipscomb. Medical Subject Headings (MeSH). *Bulletin of the Medical Library Association*, 88(3):265–6, July 2000.
7. G. A. Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.
8. D. Naber. German full form lexicon.
9. D. Naber. jWordSplitter - Java library for decomposition of German compound words.
10. J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–1543, 2014.

11. M. Sahlgren. The distributional hypothesis. *Italian Journal of Linguistics*, 20(1):33–54, 2008.
12. C. Spreckelsen, S. Finsterer, J. Cremer, and H. Schenkat. Can social semantic web techniques foster collaborative curriculum mapping in medicine? *Journal of medical Internet research*, 15(8):e169, Jan. 2013.
13. J. Steinhaeuser, J. Chenot, and M. Roos. Competence-based curriculum development for general practice in Germany: a stepwise peer-based approach instead of reinventing the wheel. *BMC Res ...*, 2013.