A Domain Specific ESA Method for Semantic Text Matching

Luca Mazzola, Patrick Siegfried, Andreas Waldis, Florian Stalder, Alexander Denzler and Michael Kaufmann

Abstract An approach to semantic text similarity matching is concept-based characterization of entities and themes that can be automatically extracted from content. This is useful to build an effective recommender system on top of similarity measures and its usage for document retrieval and ranking. In this work, our research goal is to create an expert system for education recommendation, based on skills, capabilities, areas of expertise present in someone's curriculum vitae and personal preferences. This form of semantic text matching challenge needs to take into account all the personal educational experiences (formal, informal, and on-the-job), but also work-related know-how, to create a concept based profile of the person. This will allow a reasoned matching process from CVs and career vision to descriptions of education programs. Taking inspiration from the explicit semantic analysis (ESA), we developed a domain-specific approach to semantically characterize short texts and to compare their content for semantic similarity. Thanks to an enriching and a filtering process, we transform the general purpose German Wikipedia into a domain specific model for our task. The domain is defined also through a German knowledge base or vocabulary of description for educational experiences and for job offers. Initial testing with a small set of documents demonstrated that our approach

L. Mazzola · P. Siegfried · A. Waldis · F. Stalder · A. Denzler · M. Kaufmann (✉)
School of Information Technology, Lucerne University of Applied Sciences,
6343 Rotkreuz, Switzerland
e-mail: m.kaufmann@hslu.ch

L. Mazzola
e-mail: luca.mazzola@hslu.ch

P. Siegfried
e-mail: patrick.siegfried@hslu.ch

A. Waldis
e-mail: andreas.waldis@hslu.ch

F. Stalder
e-mail: florian.stalder@hslu.ch

A. Denzler
e-mail: alexander.denzler@hslu.ch

covers the main requirements and can match semantically similar text content. This is applied in a use case and lead to the implementation of an education recommender system prototype.

Keywords Semantic text matching · Document similarity · Concept extraction · Explicit semantic analysis · Domain-specific semantic model

1 Introduction

Human consulting is expensive and time consuming. In the area of HR consulting, giving advice on possible job placements and possible further education could be automated. The vision is that users can upload their CV and their career goal, and an expert system recommends the best possible option. One of the issues for building an effective recommender system for job placement and further education programs is the difficulty of automatically identifying the skills, capabilities and areas of expertise that a person has. This is even more difficult when the person, on top of the mix of formal, informal, and on-the-job educational experiences has also work-related know-how.

In a research project partially financed by the Innovation and Technology commission (CTI) of the Swiss Confederation, we identified a possible technical solution for this problem. There is already an extensive knowledge of approaches in the state of the art, but none of the existing approaches are well tailored to our problem. In fact, the problem is characterized by the following main aspects:

(a) the need for analyzing unstructured and semi-structured documents,
(b) commitment at extracting a semantic signature for a given document,
(c) obligation to treat documents written in German since most documents in Switzerland are written in this language,
(d) usage of semantic concepts also in German,
(e) capability of running analysis on multi-parted sets finding ranked assignments for comparisons, and
(f) capacity to run with minimal human intervention towards a fully automated approach.

For these reasons, we performed research on a new approach to extract concepts and skills from text using a domain specific ESA space that is described in this work. The rest of the paper is organized as follows: Sect. 2 presents a very brief overview of related work, then our approach is described in Sect. 3 covering the different aspects of the functional requirements, the design of the system, and the data source characterization. Section 4 reports the requirement validation from the tuning of the parameters to the experimental settings. In Sect. 5, an initial evaluation with a business case oriented test bed is provided. Two use cases with different objectives

are reported in Sect. 6, demonstrating the applicability of our approach to specific instances of real problems. The conclusions (Sect. 7) recapitulate our contribution for a solution to this problem stressing also some future work we intend to address in the next step of this research project.

2 Related Work

Our proposed solution was inspired by numerous previously existing approaches and systems. For example, in the domain of document indexing, comparison and most similar retrieval there is a good review in the work of Alvarez and Bast [1], in particular with respect to word embedding and document similarity computations. Another very influential article by Egozi et al. [2], on top of supporting a concept-based information retrieval pathway, provided us with the idea of the map model called ESA (explicit semantic analysis) and also suggested some measures and metrics for the implementation. A following work by Song and Roth [3] suggested the idea of filtering the model matrix and the internal approach for sparse vector densification towards similarity computation whenever we have as input a short text. The idea of kicking-off from the best crowd-based information source, Wikipedia, was supported by the work of Gabrilovich and Markovitch [4], who described their approach for Computing semantic relatedness using wikipedia-based explicit semantic analysis. This also fits our need of a German-specific knowledge base, as wikipedia publicly provides separated dumps for each language. Recently, a work from two *LinkedIn* employees [5] showed a different approach to map together profiles and jobs with perceived good matches by using a two step approach for text comprehension: relying on the set of skills S existing on the users profiles, the job description is mapped by a neural network (Long Short Term Memory) into an implicit vectorial space and then transformed into an explicit set of related skills $\in S$ using a linear transformation of multiplicative matrix $\boldsymbol{W}$. Since embedding is a key feature, we also analyzed the work of Pagliardini et al. [6], which focuses on the unsupervised learning of sentence embeddings using compositional n-gram features, and we relied on one of our previous work [7] to extract the candidate concepts from the domain. Another possibility for achieving this task could have been to adopt the *embedRank* of Bennani et al. [8] in which they suggest an unsupervised key-phrase extraction using sentence embeddings. It is possible that focusing on the usage of information granulation for fuzzy logic and rough sets applications could be beneficial for this objective [9], together with its underlying contributions to interpretability [10].

3 The Approach

The general objective of this work is to design, implement and evaluate a data-based system that is able to compare the education steps and experiences of a person

(generally know as *Curriculum Vitae* or CV for short) in terms of keywords with possible education programs, and to semantically match them for recommendation. This means extracting from a CV its major points. To this objective, the initial prototype was devoted to analyze a single document, returning the extracted signature for human operator usage.

As this approach is useful for human expert direct consumption, but suboptimal for further more abstract tasks such as direct document comparisons, similarities extraction or document matching, there is a need for a novel type of solution, which is able to satisfy all the imposed requirements, specified in the next section.

3.1 Functional Requirements

Given the objective and the state of the art described, as starting point, we elicited some requirements through direct discussions with experts: employees of a business partner who do manual CV assessment and personalized suggestion of further educational steps on a daily basis. As a result of these interactions and the related iterative process of refinements, a common set of needs emerged as functional requirements useful to achieve common goals present in their day to day practice. Matching this candidate set with the business requirements expressed by the project partner, we eventually identified a core group of considerations stated in the following list:

1. develop a metric for comparing documents or short texts based on common attributes' sets
2. compare two given documents:
   - 2.1 identify similarities between two education-related documents
   - 2.2 extract the capabilities, skills, and areas of expertise common to two (or more) documents.
3. compare a given document against a set:
   - 3.1 assign the most relevant related job posting to a given CV
   - 3.2 find the closest education program to a CV based on a common skill-set
   - 3.3 find CVs similar to a given one in term of capabilities, skills, and areas of expertise.

Also, we identified some additional nice-to-have capabilities, such as: (a) the use of a granular approach [11] for semistructured documents, to improve their concept-based signature (b) the capability of using different knowledge metrics, (such as presence, direct count, count balanced against frequency and normalized count balanced against frequency) for considering the keyword occurrences into documents [12], and (c) the usage of different distance metrics (such as cosine distance/similarity and multi-dimensions euclidean distance) for comparing vector entries into the knowledge matrix, also called "*semantic distance*" measure [13].

3.2 System Design

The system is designed to create a matrix representing the relationship between sets of keywords and concepts. We define concepts, following the ESA approach [4], by using the wikipedia German version (called DEWiki in the rest of the paper). This means that we consider every page existing in this source as a concept, using as its identifier the page title and as description the text body (except the metadata part). The definition of a valid concept is in itself a research subject, and we built upon our previous work about concepts-extraction from unstructured text [7], to adopt the same approach. Figure 1 presents the two processes of *enriching* and *domain specific filtering* that constitute our pipeline to go from the source dump to the knowledge matrix.

Enriching is the process used to extract the complete set of valid pages, meaning all pages with a valid content (eg: excluding *disambiguation pages*) and also enriched by simulating an actual content for the *Redirect pages*. In particular, the filtering process eliminates a page whenever at least one of the following conditions holds:

- the title is entirely numerical (only consisting of a single number)
- the length of the title is equal to one
- it is a disambiguation page (no actual content, only pointers to the term different meanings)
- the page is associated with geo-tagging metadata
- the page text start with a redirect or a forward link.

Domain specific filtering refers to our intuition that instead of using a generic, transverse knowledge base, we would like to have a more focused and specific model, only covering the concepts relevant for our application domain. Nevertheless in order to not lose too much coverage, we allow redirected pages if at least one of its incoming links is part of our domain. This process preserves all wikipedia pages that are part of the set generated by computing all valid ngrams (from a vocabulary of education descriptions) for the domain specific texts (without including any punctuation).

After these two steps, the dataset is ready to be transformed into the knowledge source. Through the use of statistical approaches, the enriched and filtered list of wikipedia pages is transformed in a bidimensional matrix, whose dimensions are the stem[1] of the words in the page content (columns) and the page names, consider as concepts (rows). The content of the matrix in the centre of Fig. 2 represents the importance of each dimension for characterising a concept. We envisioned four different metrics to use for the creation of this space: BINARY (presence or absence of the stem in the page), *Term Frequency* (TF, sum of the number of appearances), *Term Frequency - Inverse Document Frequency* (TFIDF, the frequency scaled by the selectivity of the stem), and its variation named TFIDF-NORMALISED (with a normalisation obtained by dividing the TF-IDF value by the sum of elements in each row. to give values between 0 and 1). Eventually, we adopted the last one of them, balancing the frequency of the stem within the document (the *TF* part), its

[1] identification of the base word, by removal of derived or inflected variations.

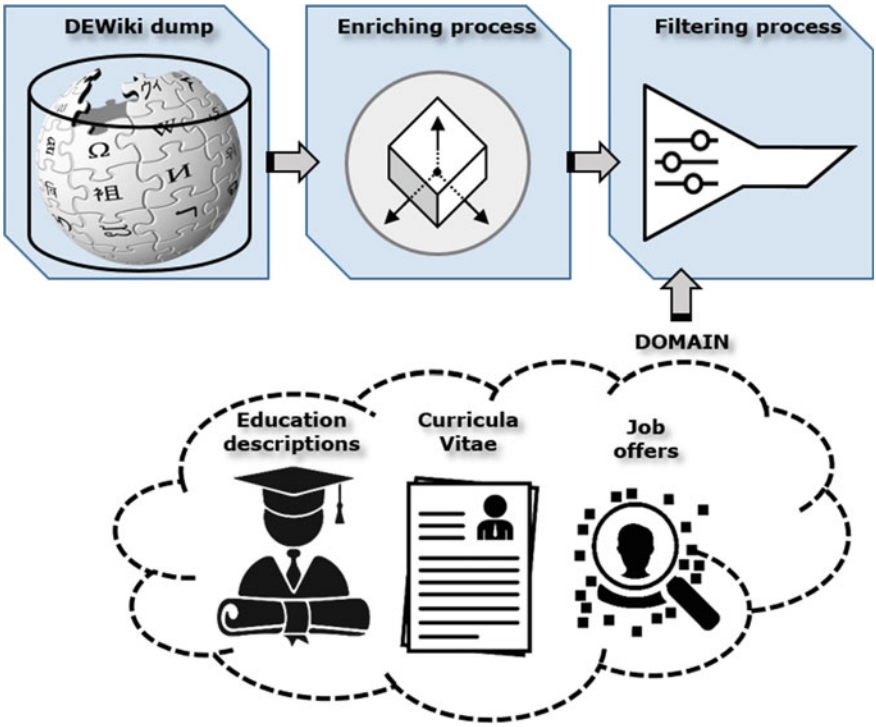

Fig. 1 The semantic matrix building process, with the two processes of *enriching* and *domain specific filtering*

specificity to the current document (as the inverse of the stem distribution amongst all the documents, the *IDF* part), and normalising the value to represent the relative importance of each single stem for the given concept.

The resulting matrix is our knowledge base, where for each wikipedia article relevant for our domain there is a distribution of stems, after filtering out too frequent and infrequent ones. Thus, every concept is represented as a vector in this knowledge space, and every short text can be transformed into such a vector and compared to the Wikipedia concepts.

It is important to note that the matrix is transposed with respect of a standard ESA model. This means that the vector space is constructed starting from stems and not wikipedia article (concepts). This difference also affects the function used for computing similarity between documents as each one of them is represented by a vector in this stems space.

Consequently, the similarity of a document to a concept can be measured by the vector distance of its stem vector to the stem vector for the concept. Accordingly, it is possible to produce a ranking of concepts for any arbitrary text document, and it

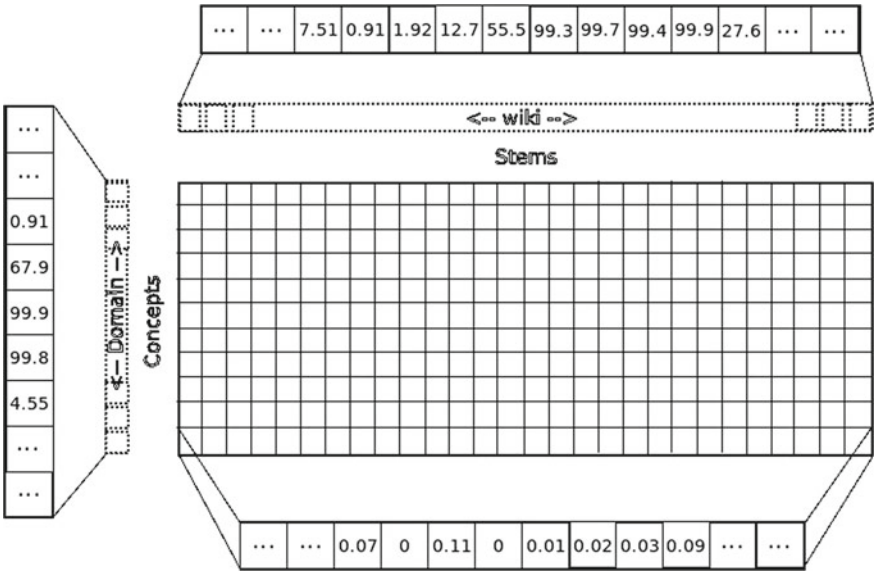

Fig. 2 The matrix and two additional data structure used to store the knowledge base for our analysis. On the rows, there are the concepts (∼800 K) known by our system derived from titles of corresponding Wikipedia entries. The columns refer to the basic stems (∼45 K) found in the full text of the Wikipedia entries for the analysis. In each cell of the matrix the weight of that component for the vector representation of the concept is stored. The two accessory multidimensional arrays maintain information about the relative position and the accumulated value of each element into the distribution, respectively for the DEWiki and the domain knowledge base. Compared to the ESA approach of Egozi et al. [2], our novel ESA matrix is transposed, having stems as dimensions to allow to position and compare not just single words in a vector space, but whole text documents as sets of words

is possible to compare the similarity of two documents by measuring the aggregated distance of their stems vectors.

As represented in Fig. 2, additional supporting data structures are maintained in order to allow restriction on the columns and rows to be taken into account for the actual computations. These consist in two bidimensional arrays that describe the relative position and the cumulated value of each element into the distribution respectively in the DEWiki and the Domain. Thanks to these supplementary information, it is possible to filter out too diffused or too specific stems and concepts, allowing a fine tune for the algorithm at run-time.

Figure 3 represents the 5 steps-long pipeline for the similarity computation for two documents, as implemented into the project demonstrator. It relies on the data structure shown in Fig. 2, here represented as the "*ESA space*".

The input documents (Doc_A and Doc_B) are parsed to extract the contained stems in *step 1* by usage of the function $stemsExtractor(Doc_x)$. This creates a ranked stem vector, using the TF measure ($\langle stems_x \rangle$). Using the domain specific and the wiki vocabularies, they are filtered in *Step 2*, as shown in the figure by the tick and x

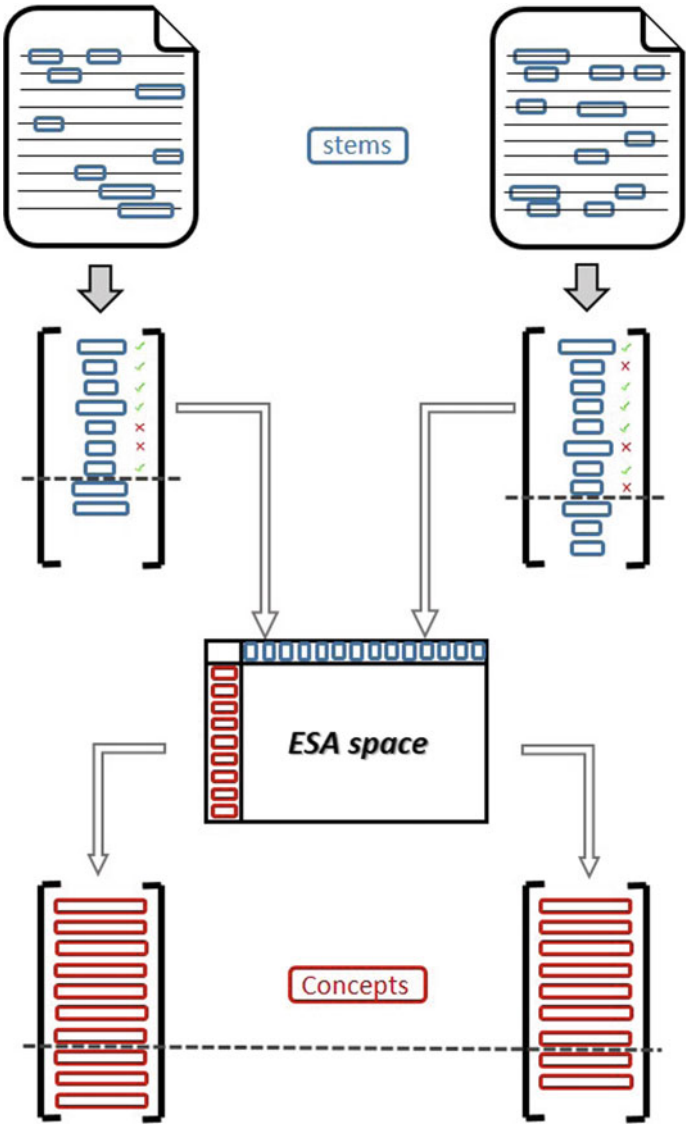

Fig. 3 The pipeline for the similarity adopted in the demonstrator is organised in 5 steps as follows. **Step 1**: starting from two documents (A and B, on the top of the figure), the stems are extracted from the document text ($stemsExtractor(X) \rightarrow \langle stems \rangle$). **Step 2**: these sets ($Stems_A$ and $Stems_B$) are then filtered, using the domain specific and the wiki vocabularies. This is the meaning of the approval or reject symbol on the side of each stem. **Step 3**: to deal with the potentially very long list of stems, and also to take into account the different length of the analysed documents, a (soft or hard) threshold is applied. **Step 4**: The resulting set of stems is then transformed into the most relevant set of concepts ($ESA(\langle stems \rangle) \rightarrow \langle concepts \rangle$), by using the calculated ESA matrix, giving $Concepts_A$ and $Concepts_B$). **Step 5**: The list of concept is compared to compute a similarity index, after a common threshold is applied to limit the input ($sim(Concepts_A, Concepts_B) \rightarrow [0, 1]$)

symbols on the side of each entry. A (soft or hard) limiting threshold is then applied on each filtered vector in *Step 3*, also in order to deal with the potentially very long list of stems, and to compensate for the potential different length of the analysed documents. The filtered and limited stems set for each document is used in *Step 4* as input for computing the relevant concepts over our calculated ESA matrix, through a mapping function ($ESA(\langle stems \rangle) \rightarrow \langle concepts \rangle$): this generates a ranked list of relevant concepts for each document ($\langle Concepts_A \rangle$ and $\langle Concepts_B \rangle$). Eventually, in *Step 5* we again limit the number of concepts (either in a "soft" approach, by accepting concepts accounting for a given percentage of the initial information, or in a "hard" way, by limiting the absolute number of concept allowed in the vector). The final result is the similarity measure in the unitary range ($Sim(\langle Concepts_A \rangle, \langle Concepts_B \rangle) \rightarrow [0, 1]$), which is computed by the weighted ratio of the common concepts over the full set of concepts.

Data Sources Characterization

The main data source is represented by a dump of the German version of wikipedia (DEWiki), taken on March 2018, and it is composed of ∼2.5 Millions pages. For the domain extension definition, we used three main data sources. The first one, composed of set of CV has a cardinality of ∼27,000, the second one, representing the description of publicly available educational experiences in Switzerland sums up to ∼1100 entries (around 300 vocational training, called "*Lehre*" in German, and 800 Higher education descriptions). The third and last source refers to open Jobs offer and has ∼30,000 postings. After enriching the initial candidate set of more then 2 millions pages, we have more than 3 millions valid entries, thanks to the removal of 253,061 irrelevant disambiguation pages and the addition of 1,443,110 "virtual" entities, derived by redirect links to 757,908 valid pages.

On this initial candidate set of pages, we apply the filtering process to restrict them only to entries relevant for our domain reducing the number of considered concepts to 39,797. To do this, we create two list of stems and their occurrences, once for the wiki and once for the domain specific documents. after that we use both of the list to filter the stems in the esa matrix. (wiki_limits and domain_limits) Consequently, the set of stems is reduced. In fact the one included in the full enriched dataset has a dimension of ∼870 K, that reduces to ∼66 K after the filtering process. These constitute the full set of dimensions.

For defining the additional data structures used in the filtering process at run-time, we computed individual and cumulated frequency of the stem and concepts in the reference model produced after the filtering process. As an example, Table 1 reports the top 10% of the distribution of the stems. In italics, the English-based stem, showing the contamination from other languages. This can be problematic since the stop-word removal and the stemming process are language dependent.

Anyway, as we demonstrate later in one experiment, it can be possible nevertheless to compare documents formulated in different documents under the condition that the domain specific vocabulary is identical. Unfortunately, in our current approach, this is not a generalised result.

Table 1 Top 10% of the stem distribution in the considered dataset

Stem	Number	Percent (%)	Cumulated (%)
gut	16,169	0.43	0.43
ch	15,870	0.42	0.86
ag	15,725	0.42	1.28
team	15,709	0.42	1.70
sowi	14,444	0.39	2.08
aufgab	13,569	0.36	2.45
bewerb	13,225	0.35	2.80
erfahr	12,880	0.34	3.15
profil	12,422	0.33	3.48
person	11,519	0.31	3.79
freu	11,422	0.31	4.09
arbeit	11,140	0.30	4.39
bereich	10,926	0.29	4.68
deutsch	10,711	0.29	4.97
such	10,523	0.28	5.25
biet	10,447	0.28	5.53
mail	10,435	0.28	5.81
of	10,352	0.28	6.09
ausbild	9668	0.26	6.34
per	9643	0.26	6.60
mitarbeit	9607	0.26	6.86
gern	9451	0.25	7.11
abgeschloss	9294	0.25	7.36
vollstand	9126	0.24	7.60
verfug	8923	0.24	7.84
kenntnis	8889	0.24	8.08
hoh	8831	0.24	8.32
kund	8454	0.23	8.54
tatig	8397	0.22	8.77
kontakt	8336	0.22	8.99
weit	8238	0.22	9.21
vorteil	8193	0.22	9.43
unterstutz	7999	0.21	9.65
berufserfahr	7813	0.21	9.85
jahr	7776	0.21	10.06

4 Implementation

To apply the document matching method described in this paper, we implemented a recommender system that matches possible education descriptions to descriptions of CVs and professional vision based on proximity in ESA space. To allow easier interaction with the demonstrator, a very simple HTML based GUI was developed profiting of the REST approach adopted in the development of the software solution as shown into Fig. 4.

The demonstrator computes the similarity amongst the (CV + Vision) text and each of the available education experience. In order to provide a fast and reactive interface, the concepts set for each available education experience is precomputed and stored instead of being computed at run-time. In this particular case, the profile used is an example for a *software developer*, whether the vision expressed the interest for extending the knowledge into the *Big Data, Machine Learning and Artificial Intelligence* direction. As result, all the proposed education experience include both aspects although in different degrees. It ranges from *Machine Learning* (both principles, practical and as element of more general Data Science approach) to specific solution for ML (*tensorFlow*), passing through *Deep Learning* and case studies.

The industry partner reportedly found these results very interesting and well aligned with what a human expert will suggest for the same input. This implicitly supports the approach, even if we still don't have any structured evaluation of the result quality.

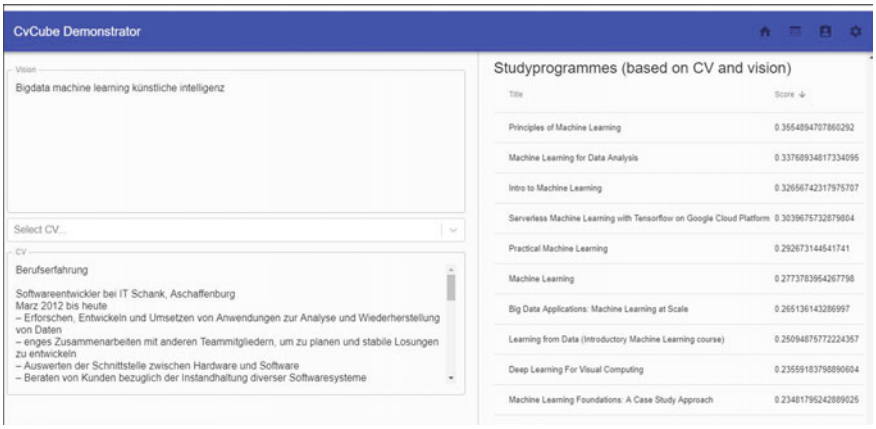

Fig. 4 A simple interface developed to allow the testing by the industry partner. The interface allows to input a *Curriculum Vitae* on the left bottom and a *Vision* text on the top of the same column. It then computes the most similar education experiences for the combination of these two elements. The column on the right reports the results, in descending order of importance

5 Evaluation

To provide the requirement (R1), we have developed a metric for comparisons of two documents. We use the balanced weight of the common concepts describing the two documents with respect to the average weight of the total set of concepts. This allows us to consider the concepts used as well as their relative pertinence to each document.

With respect to the comparison of two documents requirement (R2), we measured the capabilities of our approach based on some examples. The same is used for both the subgoals: for (R2.1) the ordered list of common concepts represent a solution, whether the consideration of the level of relevance provides an indication of the capabilities, skills and areas of expertise underlining the similarity level reported providing in this way the (R2.2) requirement.

With respect to the requirement (R3), this is a generalization of the previous category with the additional demand of considering a bigger set of documents for comparison. Despite the similarity of the internal approach required to satisfy FR3, computationally this is a more challenging problem, and we developed an additional set of functions to run, compare and rank the results of individual comparisons. Every subcategory into this requirement is distinguished by the type of resulting documents (R3.1: CV $\longmapsto$ Jobs, R3.2: CV $\longmapsto$ Education, and R3.3: CV $\longmapsto$ CVs) used for the comparison, but the algorithm to provide the results is substantially identical.

5.1 Parameters Tuning

As the system has multiple parameters to control its behavior, we ran a multi-parametrized analysis to discover the best configuration. One problem is due to the limited dimension of the test-set available since preparing the dataset and the human expert based assignment is a time consuming activity. Despite the risk of overfitting on the obtainable cases, we perceived the usefulness of this analysis.

For this, we developed a piece of code to generate a discrete variation of the set of parameters and we used these criterion lots for finding the best (most related) assignment for each document. In order to compare the result, we used a transformation matrix for generating a mono-dimensional measure from the assignment results. Table 2 presents the multipliers used. For the *top-K* documents in the ordered result set, the number of entries common with the human-proposed solution is counted and then this number is multiplied by the value present into the matrix to give one component of the global summation. In this way, we are able to directly compare runs based on different parameters set.

The set of parameters controlling our system is as follows:

- *wiki_limits*, controls the rows used by restricting too frequent or infrequent entries using the first additional multidimensional array of cumulated frequency in Fig. 2,

Table 2 Transformation for a mono-dimensional quality measure

Rank	#1	#2	#3
Top-1	2	-	-
Top-2	1/2	3/2	-
Top-3	1/3	3/3	5/3
Top-5	1/5	3/5	5/5
Top-10	1/10	3/10	5/10

meaning computed referring to the DEWiki. It is composed by a top and a bottom filtering level.

- *domain_limits*, also controls the rows to be considered in the computation, based on the cummulated frequencies into the Domain corpus. It is based on the second additional multidimensional array in Fig. 2.
- *top_stems*, indicated the maximum number of vector components that can be used to characterize at run-time a concept. It dynamically restrict the columns considerable for comparisons, by ranked absolute filtering.
- *concept_limitation_method*, controls the way concepts limitation is done: it assumes a value in the set {*HARD*,*SOFT*}. In the first case instructs the system to use an absolute number, whether in the second to conserve a certain information percentage. The value to use is respectively given by the following parameters:
  - *top_concepts* is the absolute number of top ranked concepts to use, normally between 25 and 1000.
  - *top_soft_concepts* is the cumulated information percentage that the considered top ranked concepts hold. It normally ranges between 0.05 and 0.30.
- *matrix_method*, is the method used to compute each cell value in Fig. 2. Currently we implemented an initial set {BINARY, TF, TFIDF, TFIDF_NORMALIZED}. For the current publication experiments we adopted the last value.
- *comparing_method*, is the method used for measuring the distance of elements (dissimilarity) in the restricted vector space between two or more documents. Currently we implemented only a metric that represent the cosine distance (COSINE).

Additionally to these parameters that affect the algorithm behavior, we have some config voices that only affect the presentation of results. The main ones amongst them are:

- *poss_level*, instructs the system on which final value to consider as a similarity threshold for indication of uncertain (under the given value) and possible (over it) similarity level. Usually set to 0.10.
- *prob_level*, indicates the dual threshold to distinguish between possible (under it) and probable (over it) similar documents. One candidate value from our experiment seems to be 0.25.

- *debug*, control the amount of information about the computation problem that the algorithm emits. It can be one of {True,False}

5.2 Demonstration of Semantic Relatedness

To demonstrate that our solution is producing semantically related results, we created a test case composed of 17 CVs and 44 different educational experience description, indicated by the business partner. As preparation, they also provided us with the three best assignments, as the golden standard. We then ran multiple bipartite analysis with different parameters sets, creating ranked association sets and measured their quality, based on the weight presented in Table 2.

The reference is the expected quality value for a purely random distribution without repetition of 44 elements for the considered top-k sets, with expected value $\mathbb{E}[\overline{Q}] \approx 0.32$.

On our set of 27 different runs we observed a quality in the range [3.96–10.39] with an average $\overline{Q} \approx 6.62$ and a dispersion measured with standard deviation of $\sigma[\overline{Q}] \approx 1.68$. This support our hypothesis that our approach (the model and its usage in the system) provides some knowledge.

Additionally, an human-based evaluation was performed, as we would like to have an estimation of the utility and effectiveness of our approach to support human reasoning. An expert from the business domain ranked five selected entries. We selected one entry we considered very successful (CV_9), one with intermediate results (CV_{11}), and three elements with not too good assignments (one with at least one match into the top-10 and two without anyone).

For the analytical data (matches and relevant score based on Table 2) we point the reader to Table 3. Here the second, third and forth columns represent the descending ordered position of the matches in the candidate list, whether the fifth column encode the quality score (Q) achieved by that configuration. Eventually, the seventh and last column provides the evaluation assigned by the human expert to the specific choices arrangement, here called *Stars* for analogy with a rating system.

Table 3 The manual evaluation of an initial test case subset. For everyone of the 5 CV, the 3 proposed assignments are evaluated against their position in the ex-ante human ranking. The last column presents the evaluation attached ex-post to this assignments sequence by the same human expert

CV ID	Opt #1	Opt #2	Opt #3	Quality	Stars
CV_3	>10	>10	>10	0	3
CV_6	>10	>10	>10	0	2
CV_9	1	2	5	6.7	4
CV_{11}	5	6	>10	0.6	2
CV_{16}	10	>10	>10	0.1	1

The range is [0–4], with highest value representing better option distribution. The selected set of five CV achieve an average value of 2.4, with values ranging from 1 to 4. For a very initial analysis of the rates given, is possible to note a high correlation of our quality measure with the stars-based expert rate. Interestingly and in contrast with the expectation, the two worst cases for our quality measure are rated with 2 and 3, indicating a nevertheless acceptable to good utility for the human judgment: we currently do have not clear explanations for this fact, and we need more experimental result to test any hypothesis.

6 Use Case

6.1 The Initial Testing

After the quantitative and qualitative evaluation of semantic relatedness, we identified an initial small set of documents to be used for running an experimental use case in semantic text matching. They are as follows:

- **Doc**$_1$: Description of the federal capacity certificate for car mechatronics engineer [*Automobil Mechatroniker EZF*]
- **Doc**$_2$: Job offer for a Software developer [*Software Entwickler*]
- **Doc**$_3$: Description of the Bachelor of Sciences in Medical Informatics ad at the BernerFachhochshule [*Bcs. MedizinInformatiker/in BFH*]
- **Doc**$_4$: Job offer for a car mechatronics specialist [*Automechatroniker @ Renault dealer*]
- **Doc**$_5$: Research group "Data Intelligence Team" at the HSLU - School of Information Technology
- **Doc**$_6$: Job offer as a general purpose Nurse [*Dipl. Pflegefachperson HF/FH 80–100% (Privatabteilung)*]
- **Doc**$_7$: Description of the general information of the Lucerne cantonal hospital on the website [*Luzerner Kantonspital*]
- **Doc**$_8$: The page "about us" of the Zug cantonal hospital website [*Zuger Kantonspital*]
- **Doc**$_9$: the news on the portal 20Minuten (http://www.20min.ch) about the technical issues VISA experienced in Europe on 01 June 2018 [*Visa hat technische Probleme in ganz Europa*]
- **Doc**$_{10}$: the news on the portal 20Minuten about the acquisition of Monsanto by Bayer on 07 June 2018 [*Bayer übernimmt Monsanto für 63 Milliarden*]

The set of ten documents was designed to have some clear correlations, but also to test the performance of the system on general purposes records such as the last two entries (*news*).

Within every document we extracted a weighted sequence of the top K concepts, which we considered as its semantic signature. The summarized result of the compu-

tation is shown on Table 4, where each cell represents the similarity measure between a couple of document in the selected set.

To support this interpretation, we compute the differentials with respect to each row using the relative similarity measures from Table 4 following the formula: $\overline{V_y} = \sum_x V_{xy}$ (coherently, the same is valid for the column, based on the formula $\overline{V_x} = \sum_y V_{xy}$), giving us the two transposed matrices. These matrices, encode the relative distance of each other document from the average ones. One of them is represented in Table 5, but we skipped it to represent the transposed ones. In this table, the different

Table 4 The similarity measure (cosine distance of stem vectors) amongst all the 10 documents in the test-case. Diagonals are not considered as they would always achieve the maximal score (1). Bigger values represent higher semantic signature similarities for the two documents affected. The last elements (line and column) represent the averages, respectively for row and column

Score	Doc_1	Doc_2	Doc_3	Doc_4	Doc_5	Doc_6	Doc_7	Doc_8	Doc_9	Doc_{10}	$\overline{V_y}$
Doc_1	–	0.160	0.153	0.478	0.106	0.202	0.117	0.146	0.114	0.174	0.183
Doc_2	0.160	–	0.285	0.227	0.341	0.157	0.183	0.269	0.238	0.213	0.230
Doc_3	0.153	0.285	–	0.186	0.235	0.369	0.360	0.367	0.265	0.176	0.266
Doc_4	0.478	0.227	0.186	–	0.201	0.144	0.183	0.231	0.233	0.342	0.247
Doc_5	0.106	0.341	0.235	0.201	–	0.126	0.178	0.258	0.252	0.200	0.211
Doc_6	0.202	0.157	0.369	0.144	0.126	–	0.432	0.42	0.221	0.148	0.247
Doc_7	0.117	0.183	0.360	0.183	0.178	0.432	–	0.447	0.283	0.201	0.266
Doc_8	0.146	0.269	0.367	0.231	0.258	0.420	0.447	–	0.345	0.262	0.305
Doc_9	0.114	0.238	0.265	0.233	0.252	0.221	0.283	0.345	–	0.302	0.250
Doc_{10}	0.174	0.213	0.176	0.342	0.20	0.148	0.201	0.262	0.302	–	0.224
$\overline{V_x}$	0.183	0.230	0.266	0.247	0.211	0.247	0.266	0.305	0.250	0.224	–

Table 5 The differential of each similarity value from Table 4 with respect to the row average: Delta$_1$: $\Delta_{xy_1} = V_{xy} - \overline{V_y} = V_{xy} - \sum_x V_{xy}$

Δ_{or}	Doc_1	Doc_2	Doc_3	Doc_4	Doc_5	Doc_6	Doc_7	Doc_8	Doc_9	Doc_{10}
Doc_1	–	–0.023	–0.030	0.295	–0.077	0.019	–0.066	–0.037	–0.069	–0.009
Doc_2	–0.070	–	0.055	–0.003	0.111	–0.073	–0.047	0.039	0.008	–0.017
Doc_3	–0.113	0.019	–	–0.080	–0.031	0.103	0.094	0.101	–0.001	–0.090
Doc_4	0.231	–0.020	–0.061	–	–0.046	–0.103	–0.064	–0.016	–0.014	0.095
Doc_5	–0.105	0.130	0.024	–0.010	–	–0.085	–0.033	0.047	0.041	–0.011
Doc_6	–0.045	–0.090	0.122	–0.103	–0.121	–	0.185	0.173	–0.026	–0.099
Doc_7	–0.148	–0.082	0.095	–0.082	–0.087	0.167	–	0.182	0.018	–0.064
Doc_8	–0.159	–0.036	0.062	–0.074	–0.047	0.115	0.142	–	0.040	–0.043
Doc_9	–0.136	–0.012	0.015	–0.017	0.002	–0.029	0.033	0.095	–	0.052
Doc_{10}	–0.050	–0.011	–0.048	0.118	–0.024	–0.076	–0.023	0.038	0.078	–
STD	± 0.074	± 0.040	± 0.051	± 0.090	± 0.044	± 0.086	± 0.079	± 0.061	± 0.032	± 0.047

Table 6 The final result of our experiment over the designed test-case with 10 documents: based on the simple summation of values in Table 5 and its transposed ($R_{xy} = \Delta_{xy_1} + \Delta_{xy_2} = \Delta_{xy_1} + \Delta_{yx_1}$), the final R measure is computed. The final similarity level is encoded by the different gradations of red. Higher saturation suggest a semantic closeness

R	Doc_1	Doc_2	Doc_3	Doc_4	Doc_5	Doc_6	Doc_7	Doc_8	Doc_9	Doc_{10}
Doc_1	–	–0.094	–0.144	**0.525**	–0.182	–0.026	–0.214	–0.196	–0.206	–0.060
Doc_2	–0.094	–	0.073	–0.024	**0.241**	–0.163	–0.129	0.003	–0.005	–0.029
Doc_3	–0.144	0.073	–	–0.141	–0.007	0.225	0.189	0.163	0.013	–0.138
Doc_4	**0.525**	–0.024	–0.141	–	–0.056	–0.206	–0.146	–0.090	–0.032	**0.213**
Doc_5	–0.182	**0.241**	–0.007	–0.056	–	–0.205	–0.120	0.000	0.043	–0.035
Doc_6	–0.026	–0.163	**0.225**	–0.206	–0.205	–	**0.353**	0.288	–0.055	–0.175
Doc_7	–0.214	–0.129	0.189	–0.146	–0.120	**0.353**	–	**0.324**	0.051	–0.087
Doc_8	–0.196	0.003	0.163	–0.090	0.000	0.288	0.324	–	**0.135**	–0.005
Doc_9	–0.206	–0.005	0.013	–0.032	0.043	–0.055	0.051	0.135	–	0.129
Doc_{10}	–0.060	–0.029	–0.138	0.213	–0.035	–0.175	–0.087	–0.005	0.129	–
Best	Doc_4	Doc_5	Doc_6	Doc_1	Doc_2	Doc_7	Doc_6	Doc_7	Doc_8	Doc_4
AVG	–0.066	–0.014	0.026	0.005	–0.036	0.004	0.024	0.069	0.008	–0.021
STD	± 0.142	± 0.082	± 0.121	± 0.162	± 0.093	± 0.190	± 0.182	± 0.141	± 0.073	± 0.089

gradation of yellow in the standard deviation bottom filed, represents the polarization of the set of result for each given entry in the set. Higher measures in this field intuitively suggest a better comprehension and differentiation of the peculiarities of a specific element with respect of the others in the set.

For a global view, we (cell-wise) summed-up the symmetric elements creating the final object represented into Table 6. For example $\boldsymbol{R} : Doc_{2\_5}$ and $\boldsymbol{R} : Doc_{5\_2}$ are both filled with the sum of $\Delta_{or} : Doc_{2\_5} = 0.111$ and $\Delta_{or} : Doc_{5\_2} = 0.130$ giving a value of **0.241**.

In this matrix the most significant similarity indications are highlighted with a red background, whose tone intensity positively correlates with their strength while considering the average and standard deviation of all the delta-based similarity metric reported for the specific document. The 11th row, represents for each column (document) the best candidate for semantic matching. The highlighting color used here indicate the "natural" clusters that emerge by the document thematic matching process. It is interesting to note that based on the fact the tint of the highlighting is defined on a column-based analysis, the same value can present different intensity, such as for $\boldsymbol{W_{3\_6}}$ and $\boldsymbol{W_{6\_3}}$.

6.1.1 Considerations on the Initial Testing

From the analysis of the results, we believe we can clearly identify some strong similarities, roughly corresponding with the darkest red-highlighted cells in Table 6:

- Doc_1 and Doc_4 are very similar, as they both describe the profession of car mechatronics engineer, even though from two different points of view (the first as a capacity certificate, whether the latter one as a job offer),

- Doc_2 and Doc_5 are quite similar, as they are both strictly related to computer science subareas: one presenting a software developer vacancy in a well-known online job platform, the other characterizing the research topics and projects carried out in the "Data Intelligence" team at HSLU-Informatik,
- Doc_3 is fairly comparable to Doc_6, as they partially reproduce the first case (even if in this case the domain is health-related); here a good case is represented by the similarity also with Doc_7 and Doc_8, that describe hospital profiles and offers.
- Doc_6, Doc_7, and Doc_8 constitute a reasonably related cluster, as they all are about health aspects and operations/service offered in the health domain. Here again, the relative relatedness of Doc_4 is present.

Eventually, Doc_9 and Doc_{10}, which are not specific of the domain used for building the system model, are included into the evaluation to showcase the effect of noise: no clear similarity emerges, but the effects of similar structure and common delimiter elements take a preponderant role, suggesting a similarity amongst each another, as also shown into Fig. 5.

6.2 An Additional Experiment

For this experiment, we retrieved existing jobs and courses descriptions, from online sources and used them without any preprocessing stage. This ensured the minimum amount of overfitting versus our corpus and our methodology of the test cases. We offer in this paper the description and the output of two main cases: the first one with two course descriptions in German for two somehow related professional areas (but usually adopting different vocabularies), and the second one with a very short professional outline in German and a longer job opening in French. In this second case, the capability of our approach to rely on the domain specific terminology is supported by the very specialist area the openings refers to.

We could not present any example using CVs, due to the private nature of the information there present and the new EU regulation on data privacy (EUGDPR).

Comparing documents in the same language
This experiment tests for the capability of abstraction of our "*ESA space*" in terms of being able to abstract on the specific stems used (for example, different registries or writing styles/habits) in favour of the meaning conveyed by the specific choice of words.

The first document describes the official federal professional certification competencies for a "Custodian":

> DOC_1: **Hauswart/-in mit eidg. FA**
>
> Hauswart/-in mit eidg. FA Hauswartinnen und Hauswarte sind ausgewiesene Führungs- und Fachspezialisten. Für grössere bzw. komplexere Arbeiten beauftragen sie nach Rücksprache mit der vorgesetzten Stelle spezialisierte externe

> Betriebe und begleiten die Ausführung. Sie verfügen über grundlegende administrative und rechtliche Kenntnisse. Sie sind zuständig für die Umsetzung der ökologischen und sicherheitstechnischen Richtlinien. Als Bindeglied oder Vermittler zwischen Nutzern, Kunden, Mietern und Liegenschaftsbesitzern leisten Hauswartinnen und Hauswarte einen wichtigen Beitrag für die Gesellschaft. EBZ Erwachsenenbildungszentrum Solothurn-Grenchen 91

The second document illustrates (on a more abstract level) a course for a "responsible for maintenance" of real estates:

> DOC_2: **Sachbearbeiter/in Liegenschaftenunterhalt KS/HEV**
>
> Nachhaltigkeit ist eines der Schlagwörter, wenn es um Immobilien und deren Unterhalt geht. Dies erfordert ein fundiertes Verständnis für den Lebenszyklus einer Immobilie. Möchten Sie den Unterhalt Ihrer eigenen Liegenschaft optimieren oder werden Sie in Ihrem beruflichen Umfeld immer wieder mit diesem Thema konfrontiert? Interessieren Sie sich für die Bausubstanz und deren Alterung, um grössere Ersatzinvestitionen frühzeitig planen zu können? Wollen Sie beim Kauf einer Liegenschaft wissen, worauf Sie bezüglich der Bausubstanz achten müssen und die Notwendigkeit von anstehenden Unterhaltsarbeiten erkennen sowie die zu erwartenden Kosten abschätzen können? Baufachleute vermitteln Ihnen einen umfassenden Überblick über die Bauteile einer Liegenschaft und deren Lebensdauer, Bauerneuerungsaufgaben und damit verbundene vertragliche Bindungen. Zudem erhalten Sie wertvolle Tipps für die Praxis mit auf den Weg. KS Kaderschulen 201

The individual set of stems coming from the *Step 3* are given in Table 7: the only common stem between Doc_1 and Doc_2 is **gross**, here represented in bold. To transform this stems into a similarity score, we can apply the same function used on the concepts vectors ($Sim(\langle X \rangle, \langle Y \rangle) \rightarrow [0, 1]$): the retrieved score will then be **0.001494**. Instead, the third column represents the concepts extracted after *Step 4* and limited by applying the threshold (only a part of the set is represented, as it would have been too long to include the full set of 150 concepts). Computing the similarity measure for these sets creates a value of **0.261876**. This clearly shows a case where a comparison in the space of stems will provide a very low (at the limit of being non-existing) similarity between the given set of two documents, but their projection into the concept space is able to extract a higher level meaning and re-balance the similarity level computed.

Comparing documents in different languages (special case)
In the second case, we shows the capabilities of comparing two documents written in different languages, under the assumption that the domain specific vocabulary is partially language-independent.

The first document is a one-line short collection of terminology in German associated with the operation of cutting with a milling machine:

Table 7 An experiment to demonstrate the capabilities of the transformation space $\langle stem \rangle \rightarrow \langle concept \rangle$ to overcome the choices of words (stems) in favour of the underlying semantics

Doc	$\langle stem \rangle$	$\langle concept \rangle$
1	[‘richtlini’,‘fuhrung’,‘gesellschaft’,‘wichtig’,‘beitrag’,‘verfug’,‘bzw’,‘nutz’, ‘solothurn’,‘vorgesetzt’,‘fa’,‘bindeglied’,‘extern’,‘spezialisiert’,‘grench’,‘stell’, ‘ausgewies’,**gross**’,‘rucksprach’,‘sicherheitstechn’,‘beauftrag’,‘fachspezialist’,‘leist’, ‘vermittl’,‘miet’,‘arbeit’,‘kenntnis’,‘kund’,‘zustand’,‘umsetz’,‘okolog’,‘komplex’, ‘rechtlich’,‘administrativ’,‘betrieb’,‘eidg’,‘grundleg’,‘ausfuhr’,‘hauswart’,‘begleit’]	[‘Religion’,‘Verordnung (EG) Nr. 1907/2006 (REACH)’,...,‘Infrastrukturmanagement’,‘Fachlaufbahn’, ’Bankbetriebslehre’,‘**Due-Diligence-Prüfung**’, ‘Bildungsberatung’,..., ‘**Facilitymanagement**’, ‘Concierge’, ‘Motivation’, ‘Versicherer’, ‘Schulleitung’, ‘**Personalentwicklung**’, ‘Teambildung’,..., ‘**Designmanagement**’, ‘Lean Management’,... ‘**Immobilienmarkt**’,..., ‘Projektkommunikation’, ‘**Energieeinsparung**’, ‘Schweizerischer Städteverband’,..., ‘Apotheke’, ‘**Bauherrenberatung**’,..., ‘ZEWO’, ‘Mediation’, ‘**Leasing**’, ‘Der Process’,...]
2	[‘worauf’, ‘**gross**’, ‘kauf’, ‘erhalt’, ‘kaderschul’, ‘vermitteln’, ‘fruhzeit’, ‘sowi’, ‘nachhalt’, ‘praxis’, ‘vertrag’, ‘liegenschaft’, ‘optimi’, ‘notwend’, ‘verbund’, ‘zud’, ‘beim’, ‘kost’, ‘baufachleut’, ‘imm’, ‘beruf’, ‘umfass’, ‘erfordert’, ‘unterhalt’, ‘der’, ‘acht’, ‘unterhaltsarbeit’, ‘interessi’, ‘bauteil’, ‘schlagwort’, ‘fundiert’, ‘bindung’, ‘erwart’, ‘umfeld’, ‘sachbearbeiterin’, ‘thema’, ‘ansteh’, ‘mocht’, ‘bausubstanz’, ‘konfrontiert’, ‘verstandnis’, ‘lebensdau’, ‘bezug’, ‘abschatz’, ‘ks’, ‘muss’, ‘lebenszyklus’, ‘erkenn’, ‘immobili’, ‘wertvoll’, ‘wiss’, ‘tipps’, ‘eig’, ‘uberblick’, ‘plan’, ‘geht’]	[‘Alkoholkrankheit’, ‘Bindella’, ‘Seele’,..., ‘**Leasing**’, ‘Geldwäsche’, ‘Social Media’, ‘Immobilientreuhänder’,..., ‘**Designmanagement**’, ‘Partner Privatbank Zürich’, ‘Denkmalpflege’, ‘Bilanz’,..., ‘Grünes Gebäude’, ‘Hermeneutik’, ‘**Immobilienmarkt**’, ‘Immobilie’,..., ‘Energiemanagement’, ‘Depression’, ‘Angela Merkel’, ‘Star Trek: Der erste Kontakt’, ‘Tierversuch’, ‘Ethik’, ‘Immobilienmakler’, ‘Coaching’, ‘Design’, ‘**Bauherrenberatung**’,..., ‘Netzwerk-Marketing’, ‘Werbung’, ‘Ganztagsschule’, ‘**Facilitymanagement**’, ‘Haus- und Familienarbeit’,..., ‘Wohnungsbau’, ‘Experiments’, ‘Behinderung’, ‘**Personalentwicklung**’, ‘**Due-Diligence-Prüfung**’, ‘**Energieeinsparung**’, ‘Rehaklinik Hasliberg’, ‘Bauökonomie’,...]

> DOC_3: **CNC Dreher**
>
> cnc dreher cnc turner cnc dreher cnc fräsercnc dreher

Instead, the other document in this set describes a job offer for an "milling machine with automatic control operator", in particular for the task of setting up and regulating them, and it is formulated in French:

> DOC_4: **Régleur CNC**
>
> régleur cnc ok job sa cnc machine operator régleur cnc l'un de nos clients, une entreprise du jura bernois active dans le développement de systèmes automatisés, cherche de manière temporaire pour renforcer son équipe un/e :régleur cnc h/fvotre mission:vous êtes en charge de la préparation du travail et des mises en train de machines de transfert cnc. vous garantissez le suivi de production de composants horlogers tout en contrôlant la qualité de celles-ci à l'aide d'outils de contrôle nouvelle génération.votre profil:·vous avez effectué un cfc de mécanicien de précision ou équivalent·vous avez de bonnes connaissances de la programmation et du réglage sur machines cnc· vous avez idéalement de l'expérience dans l'usinage de composants horlogers·vous êtes autonome, précis, polyvalent et curieuxintéressé(e)? dans ce cas frédéric maugeon se réjouit à l'avance de la réception de votre dossier complet. merci de transmettre votre candidature en cliquant sur "postuler"

The individual set of stems coming from the *Step 3* are given in Table 8: the only common stem between Doc_3 and Doc_4 is **cnc**, again represented in bold. Anyway, due to the minimal number of stems in the Doc_3 vector, this single element is already able to produce a similarity measure not close to zero, namely having the value of **0.208347**. Given the specificity of the vocabulary adopted by these two documents, their projection into the semantic (ESA) space is able to stress the similarity of the underlying concepts, producing a value for the similarity in *Step 5* of **0.771173**. The fact that many of the concepts retrieved that are common to this set are indeed strictly related to the milling machine world is noteworthy. They are definitively more semantically oriented than in the previous example, which included a significant portion of more generic concepts in the vectors intersection.

7 Conclusions

In this work, we presented an ESA-inspired, domain-specific approach to semantically characterizing documents and comparing them for similarities. After clarifying the usage context and the functional requirements, we described the creation of a model that sits at the core of our proposal. The peculiarities of our approach are the enriching and filtering processes,, which allow the starting from a general purpose

Table 8 The following is an experiment to demonstrate the capabilities of the computing similarities in documents about very specialised domains, regardless of the language in which they are formulated

Doc	$\langle stem \rangle$	$\langle concept \rangle$
3	[‘**cnc**’,‘turn’,‘dreh’]	[‘**Schraube**’,...,‘Palettenwechsler’, ‘Steuerungstechnik’,‘**Polymechaniker**’, ‘Rundschleifmaschine’,...,‘IEEE 1284’,‘**Schnelle Produktentwicklung**’,‘Kinästhetik’, ‘Präzision’,..., ‘Roto Frank’, ‘**Drehen (Verfahren)**’, ‘**Fanuc**’,..., ‘Drehbank’, ‘Drechsler’, ‘Tischler’, ‘**Zerspanungsmechaniker**’, ‘*CNC-Fachkraft*’, ‘**CAD**’, ‘**Werkzeugmechaniker**’,..., ‘CNC-Drehmaschine’, ‘**Häner**’, ‘**Fitting**’, ‘AutoCAD’,..., ‘**DMG Mori K.K.**’,..., ‘**Werkzeugmaschinenfabrik Arno Krebs**’, ‘Sinumerik’, ‘Feldbus’, ‘Arbeitsumgebung’, ‘CNC-Maschine’,..., ‘**Produktionswirtschaft**’,..., ‘**Drehmaschine**’, ‘**Metallurgie**’, ‘Formenbau’,...]
4	[‘job’, ‘train’, ‘cas’, ‘production’, ‘client’, ‘outil’, ‘contrôl’, ‘ci’, ‘jura’, ‘développement’, ‘travail’, ‘équip’, ‘effectué’, ‘aid’, ‘**cnc**’, ‘précis’, ‘précision’, ‘cfc’, ‘avanc’, ‘nouvell’, ‘autonom’, ‘réception’, ‘machin’, ‘charg’, ‘régleur’, ‘bernois’, ‘manièr’, ‘programmation’, ‘mis’, ‘ok’, ‘mécanici’, ‘bonn’, ‘qualité’, ‘merci’, ‘candidatur’, ‘cherch’, ‘dossi’, ‘cell’, ‘tout’, ‘temporair’, ‘horlog’, ‘frédéric’, ‘polyvalent’, ‘complet’, ‘connaissanc’, ‘entrepris’, ‘activ’, ‘système’, ‘operator’, ‘suivi’, ‘préparation’]	[‘Werkzeugschleifen’, ‘Swatch Group’, ‘Cadwork’, ‘Landwirtschaftsschule’, ‘Arbeitsumgebung’,..., ‘**Zerspanungsmechaniker**’,..., ‘**CAD**’, ‘*CNC-Fachkraft*’,..., ‘**Fitting**’,..., ‘**Schraube**’,..., ‘**Häner**’, ‘**Polymechaniker**’, ‘Bearbeitungszentrum’,..., ‘**Metallurgie**’, ‘Formenbau’, ‘Prototypes’, ‘Thermografie’, ‘Präzision’, ‘**Werkzeugmaschinenfabrik Arno Krebs**’, ‘Machine to Machine’,..., ‘**Fanuc**’,..., ‘Produktionstechnik’, ‘**DMG Mori K.K.**’, ‘LNS SA’, ‘Drawing Interchange Format’, ‘Digitalisierung’, ‘**Schnelle Produktentwicklung**’, ‘Tebis’,..., ‘**Drehmaschine**’,..., ‘Computer-aided manufacturing’, ‘**Produktionswirtschaft**’, ‘Fräsmaschine’, ‘IEEE 1284’, ‘Feldbus’, ‘Maschinelles Lernen’,..., ‘**Drehen (Verfahren)**’, ‘CNC-Drehmaschine’,..., ‘AutoCAD’, ‘DMG Mori Aktiengesellschaft’,..., ‘**Werkzeugmechaniker**’, ‘Senkerodieren’,...]

corpus of documents and create a domain specific model. This computation happens at the system initialization stage, offering a model ready-to-use at run-time. To improve the performance, we designed additional data structures and parameters to allow a more fine grained adjustment for each execution. On top of the model, we designed functions and metrics to use from seamless documents characterization and similarity scoring.

The challenge of the ESA approach proposed in [2] is the aggregation of vector representation from single words to whole documents, as this is the unity in our application domain. To solve this issue, we contribute a new ESA approach with a transposed vector space consisting of stems, representing Wikipedia Text concepts as points in this space. This allows the positioning of arbitrary text documents in this space and to compare their similarities to Wikipedia entries and all other text documents using Vector distance. Our conclusion is even though this method is not directly applicable for concept extraction like traditional ESA, we have shown that our method produces meaningful results for semantic document matching based on similarity if the set of concepts similar to two texts is compared.

We applied our approach to curricula vitae, defining our domain through a German knowledge base of description for educational experiences and for job offers. We initially statistically demonstrated that the produced results are semantically related, based on a quality mono-dimensional measure transformation of the results. From this we can conclude that some semantics is captured by our approach. Furthermore we designed a small set of 10 documents for a use case, divided into 3 clusters, with 2 unrelated elements. From that similar documents were grouped by the algorithm and thus our algorithm demonstrated the potential for semantic text matching, starting from heterogeneous sources.

Through our contribution, we show that the idea of restricting the knowledge based for the ESA space to a specific domain and the possibility to filter too common or infrequent elements from both the dimensions of the model seems to improve the capability of recognizing semantic relationship amongst documents, by reducing the noise affecting the system.

Figure 5 shows the dendrogram (hierarchical tree) produced by the normalization of the distance matrix using the *complete* approach, to balance the clusters by reducing the summation of the inter-cluster distance.

The major limit of our approach is its language dependency because the model is produced on a specific language-based jargon. Unfortunately, this is currently a structural limit since we developed our model on the German language, which is the more prominent language used in Switzerland. The job offers and the educational experience are specific for Switzerland and described in the same language. We do not expect any major issues (except the potential lack of data) in repeating the full process using sources in different languages.

Currently, this prototype is being used for comparison with manually annotated CVs in order to assess its stability (absence of macroscopic false positive) and also to verify its usefulness (in term of additional enrichment it can produce with respect to the information a human operator in a typical iteration produces). No structural result is still available in this respect as the testing is still in a initial phase.

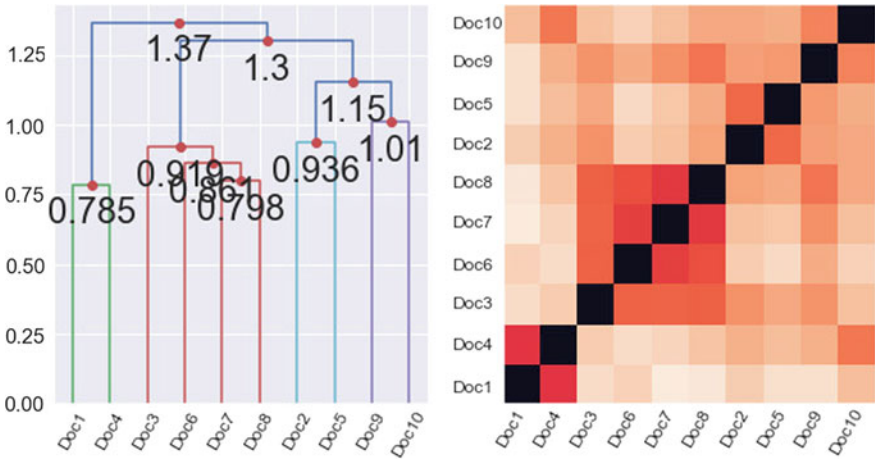

Fig. 5 *Right*: The heatmap of the document distances, for the use case in Table 6. Color saturation positively correlate with the score. *Left*: The corresponding dendrogram: here the cluster are highlighted by the use of different colours. We predefined the presence of exactly 4 clusters

Despite the promising results, we would like to improve the system and extend the testing with particular respect to:

1. implementing a quantitative benchmarking of the document matching method based on several gold standards,
2. adoption of a granular approach. We expect to improve the document characterization by its concept-based signature, in particular considering that curricula vitae are intrinsically already semistructured documents,
3. development of customizable metrics for stems weighting into the domain-specific model allowing the selection at runtime of which one to adopt for a specific run,
4. envision of different distance metrics for comparing vector entries into the knowledge matrix in order to stress distinctive aspects of our model vector space
5. estimation of the effects of parameters choice to the output, in order to identify optimal parameters sets,
6. ideate an approach to deal with multiple languages. Switzerland is a multi-lingual entity, and this will be definitely interesting, but also towards the capability of comparing documents written in different languages or to consider entries with section in various languages. An idea we are assessing is to create different ESA model, each one starting from a dump in the relevant language, and then somehow relate them using the metadata stating the equivalence of pages in different languages (normally present in Wikipedia as "*Languages*" in the bottom left of a page).

Some of these aspects will be researched in the next project steps, together with the concurrent semi-automatic creation of a lightweight ontology for concepts existing into our domain.

Acknowledgements The research leading to this work was partially financed by the KTI/Innosuisse Swiss federal agency, through a competitive call. The financed project KTI-Nr. 27104.1 is called *CVCube: digitale Aus- und Weiterbildungsberatung mittels Bildungsgraphen*. The authors would like to thank the business project partner for the fruitful discussions and for allowing us to use the examples in this publication. We would like to thank Benjamin Haymond for his very helpful and precise revision and language editing support of this manuscript.

References

1. J.E. Alvarez, H. Bast, A review of word embedding and document similarity algorithms applied to academic text, in *Bachelor's Thesis*, University of Freiburg (2017). https://pdfs.semanticscholar.org/0502/05c30069de7df8164f2e4a368e6fa2b804d9.pdf
2. O. Egozi, S. Markovitch, E. Gabrilovich, Concept-based information retrieval using explicit semantic analysis. ACM Trans. Inf. Syst. (TOIS) **29**(2), 8 (2011)
3. Y. Song, D. Roth, Unsupervised sparse vector densification for short text similarity, in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (2015), pp. 1275–1280
4. E. Gabrilovich, S. Markovitch, Computing semantic relatedness using Wikipedia-based explicit semantic analysis, in *IJcAI*, vol. 7 (2007), pp. 1606–1611
5. D. Bogdanova, M. Yazdani, *SESA: Supervised Explicit Semantic Analysis*. *arXiv preprint* arXiv:1708.03246 (2017)
6. M. Pagliardini, P. Gupta, M. Jaggi, *Unsupervised Learning of Sentence Embeddings Using Compositional n-gram Features arXiv preprint* arXiv:1703.02507 (2017)
7. A. Waldis, L. Mazzola, M. Kaufmann, Concept extraction with convolutional neural networks, in *Proceedings of the 7th International Conference on Data Science, Technology and Applications (DATA 2018)*, vol. 1 (2018), pp. 118–129
8. K. Bennani-Smires, C. Musat, M. Jaggi, A. Hossmann, M. Baeriswyl, *EmbedRank: Unsupervised Keyphrase Extraction Using Sentence Embeddings*. *arXiv preprint* arXiv:1801.04470 (2018)
9. Y. Yao et al., Granular computing: basic issues and possible solutions, in *Proceedings of the 5th Joint Conference on Information Sciences*, vol. 1 (2000), pp. 186–189
10. C. Mencar, *Theory of Fuzzy Information Granulation: Contributions to Interpretability Issues* (University of Bari, 2005), pp. 3–8
11. M.M. Gupta, R.K. Ragade, R.R. Yager, *Advances in Fuzzy Set Theory and Applications* (North-Holland Publishing Company, 1979)
12. G. Salton, C. Buckley, Term-weighting approaches in automatic text retrieval. Inf. Process. Manag. **24**(5), 513–523 (1988)
13. K. Lund, C. Burgess, Producing high-dimensional semantic spaces from lexical co-occurrence. Behav. Res. Methods Instrum. Comput. **28**(2), 203–208 (1996)