A Domain Specific ESA Inspired Approach for Document Semantic Description

Luca Mazzola[1*], Patrick Siegfried[1], Andreas Waldis[1], Michael Kaufmann[1], and Alexander Denzler[1]

[1] *HSLU - Lucerne University of Applied Sciences; School of Information Technology*, CH-6343 - Rotkreuz, Switzerland
{luca.mazzola, patrick.siegfried, andreas.waldis, m.kaufmann, alexander.denzler}@hslu.ch [*] mazzola.luca@gmail.com

Sept. 2018

Abstract

Document semantic similarity is a current research field, in particular when the concept-based characterization (or signature) of the entities should be automatically extracted from their content. This becomes critical whenever someone would like to build an effective recommender system on top of this similarity measure and its usage for document retrieval and ranking. In this work, our research goal is an expert system for job placement, based on skills, capabilities, areas of expertise present into someone's curriculum vitae and personal preferences. The challenge is 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 with existing job positions, but also towards additional educational experience for profile improvement. Taking inspiration from the explicit semantic analysis (ESA), we developed a domain-specific approach to semantically characterize documents and to compare them for similarity. Thanks to an enriching and a filtering process, we transform the general purpose German Wikipedia dump into a domain specific model for our task. The domain is defined also through a German knowledge base of description for educational experiences and for job offers. Initial testing with a small set of documents demonstrated that our approach covers the main requirements. There are still open issues that we would like to tackle in the next project steps. Alongside, we have other research directions we plan to take into account, ranging from the consideration of information granulation theories to the best parameters set for algorithm tuning, till the extensibility of our solution to a multi-lingual context.

Index terms— *Text Semantic Description*, *Documents Similarity Computation*, *Concept-Based Document Characterization*, *Explicit Semantic Analysis*, *Domain-Specific Semantic Model*

1 Introduction

One of the issue for building an effective recommender system for job placement is the difficulty of identifying the skills, capabilities, 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 finance by the Innovation and Technology commission (CTI/KTI) of the Swiss confederation, we identified a possible technical solution for this open problem. There is already a quite extensive knowledge of approaches in the state of the art, but none of the existing approach is well tailored for our problem. In fact, it is characterized by the following main aspects:

a) 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 language, as this is the most used language for documents in Switzerland,

d) usage of semantic concepts also in the German language,

e) capability of running analysis on multi-parted sets finding ranked assignments for comparisons, and

f) ability to run with minimal human intervention, towards a fully automated approach.

For these reasons, we performed research on a new approach, that is described in this work. The rest of the paper is organized as follows: Sect. II presents a very brief overview of related works, then our approach is depicted in Sect. III, covering the different aspects of the functional requirements, the design of the system, and the data source characterization. Sect. IV reports about the requirement validation, from the tuning of the parameters, till the experimental settings. Following, the results of our first experiment on a reduced dataset is presented in Sect. V, and the conclusions (Sect. VI) recaps our contribution towards a solution for this problem, stressing also some future work we intend to tackle in the next step of this research project.

2 Related Work

The solution we are proposing took inspiration by numerous already 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 us some measures and metrics for the implementation. A following work by Song and Roth [3] suggested us 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 strating from the best crowd-based information source, Wikipedia, was supported by the work of Gabrilovich and Markovitch [4], that 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 different 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 texts 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 $\boldsymbol{W}$.

As the embedding is a key feature, we also analyzed the work of Pagliardini et al. [6], towards an 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], where they suggest an unsupervised key-phrase extraction using sentence embeddings. Eventually, also the work towards the usage of information granulation for fuzzy logic and rough sets applications will be beneficial for this objective [9], together with its underlying contributions to the interpretability issues [10].

3 The Approach

The general objective of this work is to design a data-based system that is able to characterize a document summarizing the education steps and experiences of a person (generally know as *Curriculum Vitae* or CV for short) in term of keywords. 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. They are three key person from the business partner, making manual assessment of CV and personalized suggestion of next educational step, 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 kernel group of desideratum, as from the following list:

FR1) develop a metric to compare documents based on common set of attributes

FR2) compare two given documents:

- **FR2.1)** identify similarity between two educational-related documents
- **FR2.2)** extract the capabilities, skills, and areas of expertise common to two (or more) documents

FR3) compare a given document against a set:

- **FR3.1)** assign the most-related job posting to a given CV
- **FR3.2)** find the closest educational experience to a CV based on the common skill-set
- **FR3.3)** find similar CVs to a given one, in term of capabilities, skills, and areas of expertise

And 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 concept every page existing in this source, using as its identifier the page title and as description the text body (except the metadata part). The definition of valid concept is in itself a research subject, and we capitalized on our previous work about concepts extraction from unstructured text [7], adopting the same approach. Figure 1 presents the two processes of *enriching* and *domain specific filtering*, that constitutes 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 the pages with a valid content (eg: excluding the so called *disambiguation pages*) and also enriched by simulating an actual content for the *Redirect pages*.

Domain specific filtering refers to our intuition that instead of using a generic, transversal knowledge base, we would like to have a more focused and specific model, only covering the concepts relevant for our application domain.

After these two steps, the dataset is ready for being 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 center of Fig. 2 represents the importance of each dimension for characterizing a concept. As a measure, we adopted *TFIDF* (*Term Frequency - Inverse Document Frequency*), balancing the frequency of the stem within the document (the *TF* part) and its specificity to the current document (as the inverse of the stem distribution amongst all the documents, the *IDF* part).

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.

It is important to notice 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.

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

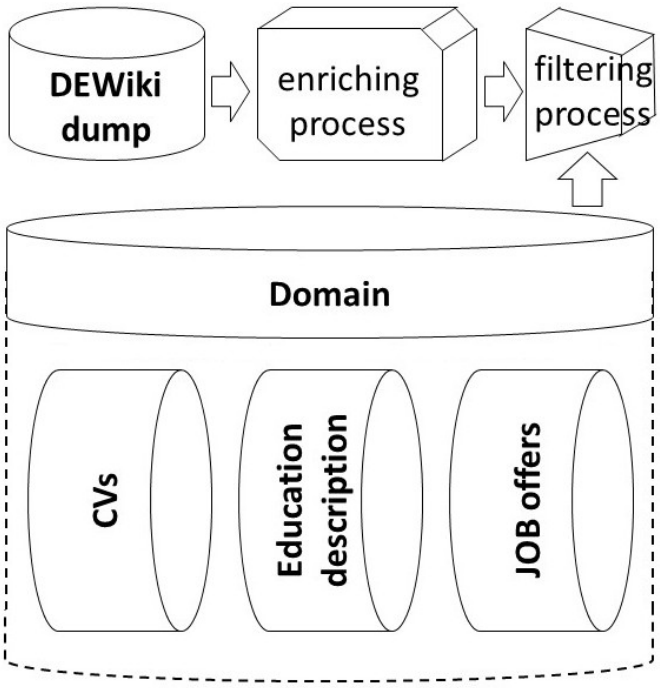

Figure 1: The semantic matrix building process, with the two processes of *enriching* and *domain specific filtering*.

Accordingly, it is possible to produce a ranking of concepts for any arbitrary text document, and it 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 on-the-flight restriction on the columns and rows to be taken into account for the actual computations. These consist in two bidimensional arrays that describes 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.

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 ex-

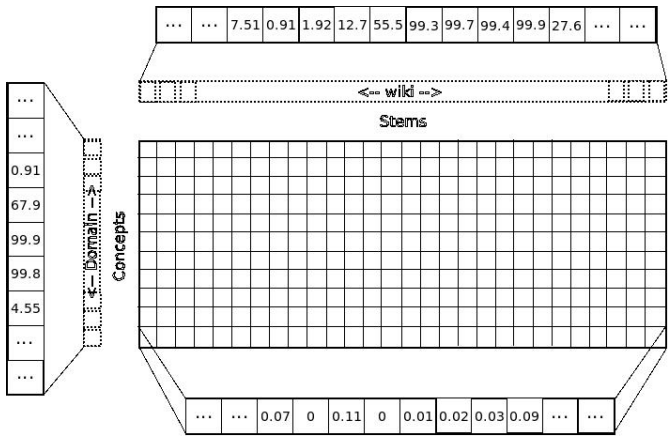

Figure 2: The matrix and two additional data structure used to store the knowledge base for our analysis. On the rows there are the concepts ($\sim$ 800K) known by our system, whether the columns refer to the basic stems ($\sim$ 45K) used for the analysis. In each cell of the matrix the weight of that component for the vector representation of the concept is stored. The two accessory multidimensional arrays maintain information about the relative position and the cumulated 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. (2011), 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.

periences in Switzerland sums up to $\sim$1.100 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 $\sim$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.

Consequently, also the set of stems reduced. In fact the one included in the full enriched dataset has a dimension of $\sim$870K, that reduces to $\sim$66K 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 lan-

guages. This can be problematic, as the stop-word removal and the stemming process are language dependent.

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

Stem	number	percent	cumulated
gut	16169	0.43%	0.43%
ch	15870	0.42%	0.86%
ag	15725	0.42%	1.28%
team	15709	0.42%	1.70%
sowi	14444	0.39%	2.08%
aufgab	13569	0.36%	2.45%
bewerb	13225	0.35%	2.80%
erfahr	12880	0.34%	3.15%
profil	12422	0.33%	3.48%
person	11519	0.31%	3.79%
freu	11422	0.31%	4.09%
arbeit	11140	0.30%	4.39%
bereich	10926	0.29%	4.68%
deutsch	10711	0.29%	4.97%
such	10523	0.28%	5.25%
biet	10447	0.28%	5.53%
mail	10435	0.28%	5.81%
of	10352	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 Requirements Validation

To provide the requirement (FR1), we 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 not only the concepts used, but also their relative relevance for each document.

With respect to the comparison of two documents requirement (FR2), we measured the capabilities of our approach based on some examples. The same is used for both the subgoals: for (FR2.1) the ordered list of common concepts represent a solution, whether the consideration also 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 (F2.2) requirement.

With respect to the requirement (FR3), 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 harder 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 (FR3.1: CV $\longmapsto$ Jobs, FR3.2: CV $\longmapsto$ Educations, and FR3.3: CV $\longmapsto$ CVs) used for the comparison, but the algorithm to provide the results is substantially identical.

4.1 Parameters tuning

As the system has multiple parameters to control its behavior, we run a multi-parametrized analysis to discover the best configuration. One problem is due to the limited dimension of the test-set available, as 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 develop a piece of code to generate 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 these way we are able to directly compare runs based on different parameters set.

4.2 Experimental Setting

The set of parameters controlling our system is as follows:

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

- *wiki_limits*, controls the rows used, by restricting too frequent or infrequent entries, using the first additional multidimensional array of cumulated frequency in Figure 2, 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 comulated frequencies into the Domain corpus. It is based on the second additional multidimensional array in Figure 2.

- *top_stems*, indicated the maximun 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 Figure 2. Currently we implemented an initial set {BINARY, TF, TF-IDF, TF-IDF_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 affects 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}

To demonstrate that our solution is not producing purely random-based set of 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 gives the evaluation assigned by the human expert to the specific choices arrangement, here called *Stars* for analogy with a rating system.

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

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

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.

5 Results

After the non purely randomness demonstration, we identified an initial small set of documents to be used for running the first experiment. 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 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 HSLU - School of Information Technologies

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 aquisition of Monsanto by Bayer on 07 June 2018 [*Bayer übernimmt Monsanto für 63 Milliarden*]

The set of 10 documents was designed to have some clear correlations, but also to test the performances of the system on general purposes records, such as the last two entries (*news*).

As for every document we extracted a weighted sequence of the top K concepts, and we considered this as its semantic signature. The summarized result of the computation is shown on Table 4, where each cell represents the similarity measure between a couple of document in the selected set.

To support the interpretation, we compute on the relative similarity measures from Table 4 the differentials with respect to each row, 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 to represent the transposed ones. In thus table, the different gradation of yellow in the standard deviation bottom filed, described how much polarized are 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.

Eventually, to have a global view, we summed-up cell-wise 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 $\mathbf{0.241}$.

In this matrix the most significant similarity indications are highlighted with a red background, whose tone intensity positively correlates with their strength, also 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, whether the highlighting color used here, indicate the "natural" clusters that emerge by the document thematic matching process. Interesting to be noted 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 $\boldsymbol{W_{6\_3}}$.

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. 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

5.1 Some Initial Considerations

From the analysis of the results, we think 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 point 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 both are related with computer sciences strictly related subareas: one presenting a software developer open position into a well-known online job platform, the other characterizing the research themes and project 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}, that 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. 3.

6 Conclusions

In this work, we presented an ESA-inspired, domain-specific approach to semantically characterize documents and to compare them for similarity. After clarifying the context of usage and the functional requirements, we described the creation of the model, that sits at the core of our proposal. Peculiarities of our approach are the enriching and filtering processes, that allows to start from a general purpose corpus of documents and create a domain specific model. This computation happens at the system initialization stage, giving a model ready to use at run-time. Anyway, to improve the performances 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 to position 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.

This will eventually fulfill the provision of an integrated solution for automatic semantic document matching in a domain oriented flavor, also towards the support of a recommender system. This expert system can then be used for job placement, based on skills, capabilities, areas of expertise present into someone's curriculum vitae and personal preferences, considering all the experiences in formal, informal, and on-the-job education, but also work-related know-how, to create a personal holistic concept based profile. It will allow a reasoned matching process with existing job positions, but also towards additional educational experience for profile improvement.

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 not random, based on a quality mono-dimensional measure transformation of the results; and then we designed a small set of 10 documents for a test-case, divided into 3 clusters, with 2 unrelated elements.

The first result was that similar documents were grouped by the algorithm (Mechanical Engineering is the common theme for (Doc_1, Doc_4) whether (Doc_2, Doc_5) refer to Computer science, the cluster composed by (Doc_6, Doc_7, Doc_8) describe health related operations/services, and so on) and thus our algorithm demonstrated the potential for documents thematic matching, starting from heterogeneous sources, even though some details still remain to be better understood. We then showcased that the main requirements are covered by the results obtained.

As our contribution, we showed 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 3 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, as the model is produced on a specific language-based jargon. Unfortunately, this is currently a structural limit, as we develop our model on the German language,

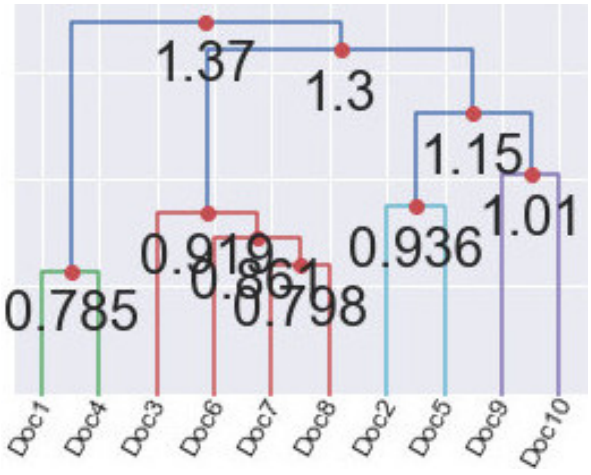

Figure 3: The dendrogram of the document distances, for the use case in Table 6. Here the cluster are highlighted by the use of different colors.

as this is the bigger language used toward Switzerland, and also the job offers and the educational experiences are specific from Switzerland and described in the same language. Anyway, we do not expect big issues (except the potential lack of data) in repeating the full process using sources in different languages.

Currently, this prototype is used for comparison with manually annotated CV, 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.

6.1 Future works

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

1. 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,

2. 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,

3. envision of different distance metrics for comparing vector entries into the knowledge matrix, in order to stress distinctive aspects of our model

vector space

4. estimation of the effects of parameters choice to the output, in order to identify optimal parameters sets,

5. ideate an approach to deal with multiple languages: as Switzerland is a multi-lingual entity, this will be definitively interesting, 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.

Acknowledgment

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.

References

[1] T. Y. Lin, "Granular computing: Fuzzy logic and rough sets." in *Computing with Words in Information/Intelligent Systems 1.* Springer, 1999, pp. 183–200.

[2] O. Egozi, S. Markovitch, and E. Gabrilovich, "Concept-based information retrieval using explicit semantic analysis." *ACM Transactions on Information Systems (TOIS)*, vol. 29, no. 2, p. 8, 2011.

[3] Y. Song and 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 and S. Markovitch, "Computing semantic relatedness using wikipedia-based explicit semantic analysis." in *IJcAI*, vol. 7, 2007, pp. 1606–1611.

[5] D. Bogdanova and M. Yazdani, "SESA: Supervised Explicit Semantic Analysis." *arXiv preprint arXiv:1708.03246*, 2017.

[6] M. Pagliardini, P. Gupta, and M. Jaggi, "Unsupervised learning of sentence embeddings using compositional n-gram features." *arXiv preprint arXiv:1703.02507*, 2017.

[7] A. Waldis, L. Mazzola, and 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, and 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*, pp. 3–8, 2005.

[11] M. M. Gupta, R. K. Ragade, and R. R. Yager, *Advances in fuzzy set theory and applications*. North-Holland Publishing Company, 1979.

[12] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval." *Information processing & management*, vol. 24, no. 5, pp. 513–523, 1988.

[13] K. Lund and C. Burgess, "Producing high-dimensional semantic spaces from lexical co-occurrence." *Behavior research methods, instruments, & computers*, vol. 28, no. 2, pp. 203–208, 1996.