Tell me why: Computational explanation of conceptual similarity judgment

Tell Me Why: Computational Explanation of Conceptual Similarity Judgments

Davide Colla, Enrico Mensa, Daniele P. Radicioni(✉), and Antonio Lieto

Dipartimento di Informatica – Università di Torino, Italy
{colla,mensa,radicion,lieto}@di.unito.it

Abstract. In this paper we introduce a system for the computation of explanations that accompany scores in the conceptual similarity task. In this setting the problem is, given a pair of concepts, to provide a score that expresses in how far the two concepts are similar. In order to explain how explanations are automatically built, we illustrate some basic features of COVER, the lexical resource that underlies our approach, and the main traits of the MeRaLi system, that computes conceptual similarity and explanations, all in one. To assess the computed explanations, we have designed a human experimentation, that provided interesting and encouraging results, which we report and discuss in depth.

Keywords: Explanation, Lexical Semantics, Natural Language semantics, conceptual similarity, lexical resources

1 Introduction

In the Information Age an ever-increasing number of text documents are being produced over time [3]; herein, the growth of the Web and the tremendous spread of social networks exert a strong pressure on Computational Linguistics to refine methods and approaches in areas such as Text Mining and Information Retrieval.

One chief feature for systems being proposed in these areas would be that of providing some sort of explanation on the ways their output was attained, so to both unveil the intermediate steps of the computation process and to justify the obtained results. Different kinds of explanation can be drawn, ranging from argumentation based approaches [23] to inferential processes triggered in formal ontologies categorisation [15]. Almost since its inception, explanation has involved expert systems and dialogue systems. In particular, the pioneering research on knowledge-based systems and decision support systems revealed that in many tasks of problem-solving "when experts disagree, it is not easy to identify the 'right answer' [...]. In such domains, the process of argumentation between experts plays a crucial role in sharing knowledge, identifying inconsistencies and focusing attention on certain areas for further examination" [21]. Explanation is thus acknowledged to be an epistemologically relevant process, and a precious feature to build robust and informative systems.

Within the broader area of Natural Language Semantics, we single out Lexical Semantics (that is, the study of word meanings and their relationships), and illustrate how COVER [18] —a resource developed in the frame of a long-standing attempt at combining ontological and commonsense reasoning [7,13]— can be used to the ends of building simple explanations that may be beneficial in the computation of conceptual similarity. COVER, so named after 'COmmonsense VEctorial Representation',[1] is a lexical resource resulting from the blend of BabelNet [22], NASARI [2] and ConceptNet [9]. As a result COVER combines, in a synthetic and cognitively grounded way, lexicographic precision and common sense aspects. We presently consider the task of automatically assessing the conceptual similarity (that is, given a pair of concepts the problem is to provide a score that expresses in how far the two concepts are similar), meantime providing an explanation to the score. This task has many relevant applications, in that identifying the proximity of concepts is helpful in various tasks, such as documents categorisation [25], conceptual categorisation [14], keywords extraction [16,4], question answering [8], text summarization [10], and many others. The knowledge representation adopted in COVER allows a uniform access to concepts via BabelNet synset IDs. The resource relies on a vector-based semantic representation which is, as shown in [12], also compliant with the Conceptual Spaces, a geometric framework for common-sense knowledge representation and reasoning [5].

In this paper we show that COVER, which is different in essence from all previously existing lexical resources, can be used to build explanations accompanying the similarity scores. To the best of our knowledge, COVER is the only lexical resource that 'natively' produces explanations: after a brief introduction of the resource (Section 2), we illustrate the main traits of the MeRaLi system, that has been designed to compute the conceptual similarity [17], and presently extended to build explanations (Section 3). We then illustrate the experimentation we conducted to assess the quality of the produced explanations (Section 4). Although the experimentation is a preliminary one, the automatic explanation has been found acceptable in many cases by the participants we interviewed. Additionally, formulating explanations seems to trigger some subtle though significant variation in the similarity judgement w.r.t. the condition in which no explanation is required, thus confirming the relevant role of the explanation in many complex tasks.

2 The COVER Lexical Resource

Let us start by quickly recalling the COVER resource [18,12]. COVER is a list of vectors, each representing and providing knowledge about a single concept. The representation of concepts rather then just terms requires the adoption of a set of concept identifiers (so to define a uniform naming space), and COVER relies on the sense inventory provided by BabelNet [22].

[1] COVER is available for download at `http://ls.di.unito.it`.

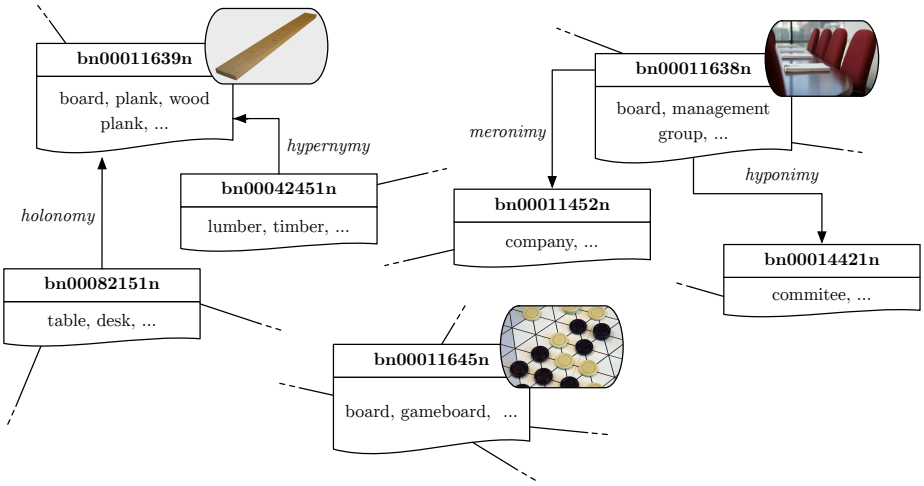

Fig. 1. A portion of BabelNet representing the possible meanings for the term *board*. Each meaning is represented as a synset, which is in turn identified uniquely by its own BabelNet synset ID. Synsets are connected via named semantic relationships.

BabelNet is a *semantic network* in which each node –called synset, that is 'set of synonyms', as originally conceived in WordNet [19]– represents a unique meaning, identified through a BabelNet synset ID (e.g., BN:00008010N). Furthermore, each node provides a list of multilingual terms that can be used in order to express that meaning. The synsets in BabelNet are also connected via a set of semantic relationships such as hyponymy, homonymy, meronymy, *etc.*. As anticipated, COVER adopts BabelNet synset identifiers in order to uniquely refer to concepts and their attached vectors. Figure 1 illustrates an excerpt of the BabelNet graph, focusing on the different meanings underlying the term *board*.

The conceptual information borrowed from BabelNet has been coupled to common-sense knowledge, that has been extracted from ConceptNet [9]. ConceptNet is a network of terms and compound words that are connected via a rich set of relationships.[2] As an example, Figure 2 shows the ConceptNet node for the term *board*.

The ConceptNet relationships have been set as the skeleton of the vectors in COVER, that is the set of D dimensions upon which a vector describes

[2] InstanceOf, RelatedTo, IsA, AtLocation, dbpedia/genre, Synonym, DerivedFrom, Causes, UsedFor, MotivatedByGoal, HasSubevent, Antonym, CapableOf, Desires, CausesDesire, PartOf, HasProperty, HasPrerequisite, MadeOf, CompoundDerivedFrom, HasFirstSubevent, dbpedia/field, dbpedia/knownFor, dbpedia/influencedBy, dbpedia/influenced, DefinedAs, HasA, MemberOf, ReceivesAction, SimilarTo, dbpedia/influenced, SymbolOf, HasContext, NotDesires, ObstructedBy, HasLastSubevent, NotUsedFor, NotCapableOf, DesireOf, NotHasProperty, CreatedBy, Attribute, Entails, LocationOfAction, LocatedNear.

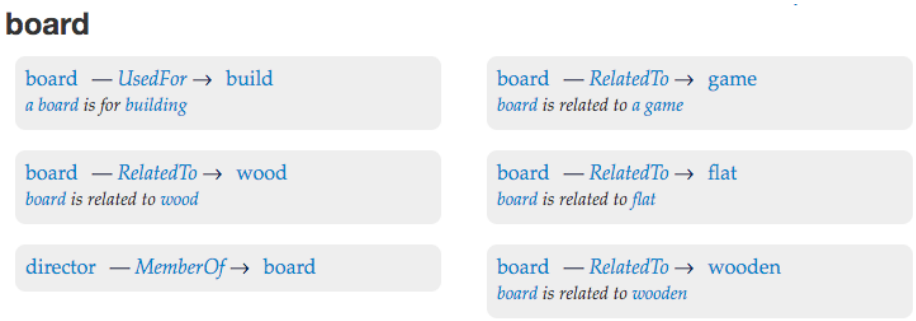

Fig. 2. Example of the ConceptNet node for the term *board*. The common-sense knowledge is encoded via a series of connections to other terms.

the represented concept. More precisely, each vector dimension contains a set of values that are concepts themselves, identified through their own BabelNet synset IDs. So a concept c_i has a vector representation $\vec{c_i}$ that is defined as

$$\vec{c_i} = [s_1^i, .., s_N^i]. \tag{1}$$

Namely, each s_h^i is the set of concepts filling the dimension $d_h \in D$. Each s can either contain an arbitrary number of values, or be empty.

For instance, the concept *headmaster* (BN:00043259N) is represented in COVER by a vector that has nine filled dimensions (RELATEDTO, ISA, HASCONTEXT, SIMILARTO, ANTONYM, DERIVEDFROM, ATLOCATION, SYNONYM, FORMOF), and therefore it has nine non-empty sets of values (Figure 3).

Exemplar BN:00043259N (head, headmaster)

$\text{BN:00043259N}_{\text{RELATEDTO}} = [\texttt{prefect}, \texttt{college}, \texttt{rector}, \texttt{teacher}, \texttt{university}, ...]$
$\text{BN:00043259N}_{\text{ISA}} = [\texttt{educator}, \texttt{head teacher}]$
$\text{BN:00043259N}_{\text{ATLOCATION}} = [\texttt{school}]$
$\text{BN:00043259N}_{\text{ANTONYM}} = [\texttt{student}]$
...

Fig. 3. A portion of the COVER vector for the *headmaster* concept. The values filling the dimensions are concepts identifiers (BabelNet synset IDs); for the sake of the readability they have been replaced with their corresponding terms.

3 Computing Conceptual Similarity

In order to compute the conceptual similarity, we designed the MeRaLi system [17]. In the conceptual similarity task, the system is provided with a pair of terms and it is required to provide a score of similarity between the two. Since the score is computed by exploiting the knowledge in COVER, one underlying assumption is that conceptual similarity can be calculated by relying on few common-sense key features that characterise the two terms at hand. More precisely, in this setting, the similarity among two terms is proportional to the amount of shared information between their respective COVER vectors.

The computation of the similarity starts with the retrieval of the proper vectors representing the terms provided as input. Terms can have multiple meanings, and therefore this search can possibly return multiple vectors for a given term. This issue is resolved by computing the similarity between all the combination of pairs of retrieved vectors, and then by choosing the highest similarity score: that is, given two terms w_1 and w_2, each with an associated list of senses $s(w_1)$ and $s(w_2)$, we compute

$$\mathrm{sim}(w_1, w_2) = \max_{\vec{c}_1 \in s(w_1), \vec{c}_2 \in s(w_2)} \left[\mathrm{sim}(\vec{c}_1, \vec{c}_2)\right]. \tag{2}$$

The similarity computation can be formally expressed as follows: given two input terms t_i and t_j, the corresponding COVER vectors $\vec{c}_i$ and $\vec{c}_j$ are retrieved. The similarity is then calculated by counting, dimension by dimension, the set of values (concepts) that $\vec{c}_i$ and $\vec{c}_j$ have in common. The scores obtained upon every dimension are then combined, thus obtaining an overall similarity score, that is our final output. So, given N dimensions in each vector, the similarity value

$$\mathrm{sim}(\vec{c}_i, \vec{c}_j)$$

is computed as

$$\mathrm{sim}(\vec{c}_i, \vec{c}_j) = \frac{1}{N} \sum_{k=1}^{N} |s_k^i \cap s_k^j|. \tag{3}$$

MeRaLi actually employs a more sophisticated formula in order to account for the possibility that the two COVER vectors may present very unequal amounts of information. Specifically, the similarity within each dimension is computed by means of the Symmetrical Tversky's Ratio Model [11], which is a symmetrical reformulation for the Tversky's ratio model [26],

$$\mathrm{sim}(\vec{c}_i, \vec{c}_j) = \frac{1}{N^*} \cdot \sum_{k=1}^{N^*} \frac{|s_k^i \cap s_k^j|}{\beta\left(\alpha a + (1-\alpha)\ b\right) + |s_k^i \cap s_k^j|} \tag{4}$$

where $|s_k^i \cap s_k^j|$ counts the number of shared concepts that are used as fillers for the dimension d_k in the concept $\vec{c}_i$ and $\vec{c}_j$, respectively; a and b are defined as $a = \min(|s_k^i - s_k^j|, |s_k^j - s_k^i|)$, $b = \max(|s_k^i - s_k^j|, |s_k^j - s_k^i|)$; finally N^* counts the dimensions actually filled with at least two concepts in both vectors. This

formula allows tuning the balance between cardinality differences (through the parameter α), and between $|s_k^i \cap s_k^j|$ and $|s_k^i - s_k^j|, |s_k^j - s_k^i|$ (through the parameter β).[3]

Example: computation of the similarity between atmosphere and ozone. As an example, we report the similarity computation between the concepts *atmosphere* and *ozone*. Firstly, the COVER resource is searched in order to find vectors

Similarity calculation for 'atmosphere' (bn:00006803n) and 'ozone' (bn:00060040n - ozone).

VDimension name	Sim	V1-V2 count	Shared	Direct	Values
InstanceOf	00.00	[000 | 000]	0	-	-
RelatedTo	00.57	[107 | 021]	8	✓	stratosphere, air, ozone, atmosphere layer, atmosphere, oxygen, gas
IsA	00.49	[004 | 005]	1	✓	gas
AtLocation	00.00	[001 | 001]	0	-	-
DBP_Genre	00.00	[000 | 000]	0	-	-
Synonym	00.00	[004 | 001]	0	-	-
DerivedFrom	00.00	[001 | 000]	0	-	-
Causes	00.00	[000 | 000]	0	-	-
UsedFor	00.00	[000 | 000]	0	-	-
MotivatedByGoal	00.00	[000 | 000]	0	-	-
HasSubevent	00.00	[000 | 000]	0	-	-
Antonym	00.00	[000 | 000]	0	-	-
CapableOf	00.00	[000 | 000]	0	-	-
Desires	00.00	[000 | 000]	0	-	-
CausesDesire	00.00	[000 | 000]	0	-	-
PartOf	00.00	[003 | 000]	0	-	-
HasProperty	00.00	[000 | 000]	0	-	-
HasPrerequisite	00.00	[000 | 000]	0	-	-
MadeOf	00.00	[000 | 000]	0	-	-
CompoundDerivedFrom	00.00	[000 | 000]	0	-	-
HasFirstSubevent	00.00	[000 | 000]	0	-	-
DBP_Field	00.00	[000 | 000]	0	-	-
DBP_KnownFor	00.00	[000 | 000]	0	-	-
influencedBy	00.00	[000 | 000]	0	-	-
DefinedAs	00.00	[000 | 000]	0	-	-
HasA	00.00	[007 | 000]	0	-	-
MemberOf	00.00	[000 | 000]	0	-	-
ReceivesAction	00.00	[000 | 000]	0	-	-
SimilarTo	00.00	[000 | 000]	0	-	-
SymbolOf	00.00	[000 | 000]	0	-	-
HasContext	00.83	[002 | 002]	1	✓	chemistry
NotDesires	00.00	[000 | 000]	0	-	-
ObstructedBy	00.00	[000 | 000]	0	-	-
HasLastSubevent	00.00	[000 | 000]	0	-	-
NotUsedFor	00.00	[000 | 000]	0	-	-
NotCapableOf	00.00	[000 | 000]	0	-	-
DesireOf	00.00	[000 | 000]	0	-	-
NotHasProperty	00.00	[000 | 000]	0	-	-
CreatedBy	00.00	[000 | 000]	0	-	-
Attribute	00.00	[000 | 000]	0	-	-
Entails	00.00	[000 | 000]	0	-	-
LocationOfAction	00.00	[000 | 000]	0	-	-
LocatedNear	00.00	[000 | 000]	0	-	-
FormOf	00.00	[001 | 000]	0	-	-

Fig. 4. Log of the comparison between the concepts *atmosphere* and *ozone* in MeRaLi. The 'V1-V2 count' column reports the number of concepts for a certain dimension in the first and second vector, respectively; the column 'Shared' indicates how many concepts are shared in the two conceptual descriptions along the same dimension; and the column 'Values' illustrates (the nominalization of) the concepts actually shared along that dimension.

[3] The parameters α and β were set to .8 and .2 for the experimentation, based on a parameter tuning performed on the RG, MC and WS-Sim datasets [17].

suitable for the representation of the two terms. The best fit resulted to be the pair of concepts ⟨BN:00006803N, BN:00060040N⟩. The similarity was then computed on a scale $[0, 1]$ by adopting Equation 4, and lately mapped onto the range $[0, 4]$. The final similarity score was 00.63, (converted to 2.52). The gold standard for this pair of terms was instead 2.58 [1]. Figure 4 illustrates the comparison table between the two vectors selected for the computation. □

Explaining Conceptual Similarity

The score of similarity provided by a system can often seem like an obscure number. It is difficult to demonstrate on which accounts two concepts are similar, especially if the score computation relies on complex networks or synthesised representations. However, thanks to the fact that COVER vectors contain explicit and human-readable knowledge, the explanation of the score is in this case allowed. Specifically, the COVER vectors adopted by the MeRaLi system provide human-readable features that are compared in order to obtain a similarity score. The explanation for this score can thus be obtained by simply reporting which values were a match in the two compared vectors. Ultimately, a simple Natural Language Generation approach has been devised: at this stage, a template is filled with the features in common between the two vectors, dimension by dimension (please refer to Figure 4). For instance, the explanation for the previously introduced example, can be directly obtained by extracting the shared values among the two considered vectors, thus obtaining:

> The similarity between *atmosphere* [bn:00006803n] and *ozone* [bn:00060040n] is 2.52 because they are *gas*; they share the same context *chemistry*; they are related to *stratosphere*, *air*, *atmosphere*, *layer*, *ozone*, *atmosphere*, *oxygen*, *gas*.

4 Experimentation

The experimentation is a preliminary pilot study, aimed at assessing the quality of the explanations. Since the language generation process itself is less relevant in the present approach, we focus on the content of the explanation rather than on the linguistic realisation.

4.1 Experimental Design

Overall 40 pairs of terms were randomly selected from the data-set designed for the shared task 'SemEval-2017 Task 2: Multilingual and Cross-lingual Semantic Word Similarity' [1] (Table 1).[4] Such pairs have been arranged into 4 questionnaires, that were administered to 33 volunteers, aged from 20 to 23. All recruited subjects were students from the Computer Science Department of the University of Turin (Italy); none of them was an English native speaker.

Questionnaires were split into 3 main sections:

[4] Actually the pair ⟨*mojito*,*mohito*⟩ was dropped in that 'mojito' was not recognised as a morphological variant of 'mohito' by most participants.

Table 1. The pairs of terms employed in each questionnaire, referred to as Q1-Q4.

Q1	desert, dune	palace, skyscraper	mojito, mohito	city center, bus	beach, coast
	videogame, pc game	medal, trainers	butterfly, rose	Wall Street, financial market	Apple, iPhone
Q2	lizard, crocodile	sculpture, statue	window, roof	agriculture, plant	flute, music
	demon, angel	income, quality of life	underwear, body	Boeing, plane	Caesar, Julius Caesar
Q3	basilica, mosaic	snowboard, skiing	pesticide, pest	level, score	snow, ice
	myth, satire	sodium chloride, salt	coach, player	Zara, leggings	Cold War, Soviet Union
Q4	car, bicycle	democracy, monarchy	pointer, slide	flag, pole	lamp, genie
	digit, number	coin, payment	surfing, water sport	Harry Potter, wizard	Mercury, Jupiter

- in the *task* 1 we asked the participants to assign a similarity score to 10 term pairs (in this setting, scores are continuous in the range $[0, 4]$, as it is customary in the international shared tasks on conceptual similarity [1]);
- in the *task* 2 we asked them to explain in how far the two terms at stake were similar, and then to indicate a new similarity score (either the same or different) to the same 10 pairs as above;
- in the *task* 3 the subjects were given the automatically computed score along with the explanation built by our system. They were requested to evaluate the explanation by expressing a score in a $[0, 10]$ Likert scale, and also to provide some comments on missing/wrong arguments, collected as open text comments.

Each volunteer compiled one questionnaire (containing 10 term pairs), which on average took 20 minutes.

4.2 Results and discussion

The focus of the present experimentation was the assessment of the automatically computed explanations (addressed in the *task* 3): MeRaLi's explanations obtained, on average, the score of 6.62 (standard deviation: 1.92). Our explanations and the scores computed automatically have been overall judged to be reasonable.

By examining the 18 pairs that obtained an averaged poor score (≤ 6), we observe that either few information was available, or it was basically wrong. Regarding the first case, we counted 12 pairs with only one or two shared concepts (please refer to Equations 3 and 4): almost always these explanations were evaluated with low scores (on average, 4.48). We found only one notable exception about the pair $\langle$*Boeing*, *plane*$\rangle$ whose explanation was

> `The similarity between` *`Boeing`* `and` *`plane`* `is 2.53 because they are related to` *`airplane`*`,` *`aircraft`*`.`

This explanation obtained an average score of 8.63. We hypothesise that this greater appreciation is due to the fact that even if only two justifications are provided, they match the most salient (based on common-sense[5] accounts) traits

[5] We refer to common-sense as to a portion of knowledge that is both widely accessible and elementary [20], and reflecting typicality traits encoded as prototypical knowledge [24].

Table 2. Correlation between the similarity scores provided by the subjects interviewed and the scores in the Gold standard. The bottom row shows the correlations between the scores gold standard and the scores computed by our system

	Spearman's ρ	Person's r
Gold - avg scores (task 1)	0.83	0.82
Gold - avg scores (task 2)	0.85	0.83
COVER - avg scores (task 1)	0.71	0.72
COVER - avg scores (task 2)	0.72	0.73
Gold - COVER	0.79	0.78

between the two considered concepts. It would seem thus that the quality of a brief explanation heavily depends on the presence of those particular and meaningful traits. In the remaining 6 pairs, vice versa, there is enough though wrong information, possibly due to the selection of the wrong meaning for input terms. In either cases, we observe that the resource still needs being improved for what pertains its coverage and the quality of the hosted information (since it is automatically built by starting from BabelNet, it contains all noise therein). This is the target of our present and future efforts.

The first and second task in the questionnaire can be thought of as providing evidence to support the result in the third one. In particular, the judgements provided by the volunteers closely approach the scores in the gold standard, as it is shown by the high (over 80%) Spearman's (ρ) and Person's (r) correlations (Figure 2). The first two rows show the average agreement between the scores *before* producing an explanation for the score itself (Gold - avg scores (*task* 1)), and *after* providing an explanation (Gold - avg scores (*task* 2)). These figures show that even human judgement can benefit from producing explanations, as the scores in *task* 2 showcase a higher correlation with the gold standard scores. Additionally, the output of the system exhibits a limited though significantly higher correlation with the similarity scores provided after trying to explain the scores themselves (COVER - avg scores (*task* 1) condition *vs.* COVER - avg scores (*task* 2)).

In order to further assess our results we also performed a qualitative analysis on some spot cases. For the pair $\langle Mercury, Jupiter \rangle$ the MeRaLi system computed a semantic similarity score of 2.29 (the gold standard score was 3.17), while the average score indicated by the participants was 3.43 (*task* 1) and 3.29 (*task* 2). First of all, this datum corroborates our approach (Section 3) that computes the similarity between the closest possible senses (please refer to Equation 2): it never happened that any participant raised doubts on the meaning of Mercury (always intended as the planet), whilst *Mercury* can be also a metallic chemical element, a Roman god, the Marvel character who can turn herself into mercury, and several further entities.

The open text comments report explanations such as that Mercury and Jupiter are *'both planets, though different'*. In this case, the participants ac-

knowledge that the two entities at stake are planets but rather different (e.g., the first one is the smallest planet in the Solar System, whilst the second one is the largest). The explanation provided by our system is:

> The similarity between Mercury and Jupiter is 2.29 because they are *planet*; they share the same context *deity*; they are semantically similar to *planet*; they are related to *planet*, *Roman_deity*, *Jupiter*, *deity*, *solar_System*.

In this case, our explanation received an average score of 9.57 out of 10. Interestingly enough, even though the participants indicated different similarity scores, they assigned a high quality score to our explanation, thus showing that it is basically reasonable.

As a second example we look at the pair ⟨*myth*, *satire*⟩. The similarity score and the related explanation of such terms are:

> The similarity between *myth* and *satire* is 0.46 because they are *aggregation*, *cosmos*, *cognitive_content*; they are semantically similar to *message*; they form *aggregation*, *division*, *message*, *cosmos*, *cognitive_content*.

In this case, the gold standard similarity value was 1.92, the average scores provided by the participants 1.57 (*task* 1) and 1.71 (*task* 2). Clearly, the explanation was not satisfactory, and it was rated 4.49 out of 10. The participants gave no clear explanation about their judgement (in *task* 2) nor informative comments/-criticisms on the explanation above (in *task* 3). One possible reason behind the poor assessment might be found in the interpretation of the *satire* term. If we consider satire as the ancient literary genre where characters are ridiculed, the explanation becomes more coherent: they are forms of *aggregation* as it was for any sort of narrative in the ancient (mostly Latin) culture; they also both deliver some message, either explaining some natural or social phenomenon and typically involving supernatural beings (like myth), or criticising people's vices, particularly in the context of contemporary politics (like satire). This possible meaning has been considered only by 2 out of 8 participants, that mostly intended satire as a generic ironic sort of text. Even in this case, where the output of MeRaLi was rather unclear and questionable, the explanation shows some sort of coherence, although not immediately sensible for human judgement. In such cases, by resorting to an inverse engineering approach, the explanation can be used to figure out which senses (underlying the terms at hand) are actually intended.

5 Conclusions

In this paper we have illustrated how COVER can be used to build explanations for the conceptual similarity task. Furthermore, we have shown that in our approach two concepts are similar insofar as they share values on the same dimension, such as when they share the same function, parts, location, synonyms,

prerequisites, and so forth; this approach is intrinsically ingrained with explanation, to such an extent that building an explanation in MeRaLi simply amounts to listing the elements actually used in the computation of the conceptual similarity score. We have then reported the experimental results obtained in a test involving human subjects over a data-set devised for an international challenge on semantic word similarity: the participants were requested to provide conceptual similarity scores, to produce explanations, and to assess the explanations computed through MeRaLi. The experimentation provided interesting and encouraging results, basically showing that when the COVER lexical resource has enough information on the concepts at hand it produces reasonable explanations. Moreover, the experimentation suggested that explanation can be beneficial also for human judgements, that tend to be more accurate (more precisely: statistically correlated to gold standard scores) after having produced explanation to justify some score in the conceptual similarity task. Such results confirm that systems for building explanations can be useful in many other semantics-related tasks, where it may be convenient (if necessary) to shepherd results and their justification.

Extending the present approach by adopting a realisation engine (such as, e.g., [6]) to improve the generation step, and devising a more extensive experimentation will be the object of our future work.

Acknowledgements

We desire to thank Simone Donetti and the Technical Staff of the Computer Science Department of the University of Turin, for their support.

References

1. Camacho-Collados, J., Pilehvar, M.T., Collier, N., Navigli, R.: Semeval-2017 task 2: Multilingual and cross-lingual semantic word similarity. In: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval 2017). pp. 15–26. Vancouver, Canada (2017)
2. Camacho-Collados, J., Pilehvar, M.T., Navigli, R.: NASARI: a novel approach to a semantically-aware representation of items. In: Procs. of NAACL. pp. 567–577 (2015)
3. Cambria, E., Speer, R., Havasi, C., Hussain, A.: Senticnet: A publicly available semantic resource for opinion mining. In: AAAI fall symposium: commonsense knowledge. vol. 10, pp. 14–18 (2010)
4. Colla, D., Mensa, E., Radicioni, D.P.: Semantic Measures for Keywords Extraction, pp. 128–140. Springer International Publishing, Cham (2017)
5. Gärdenfors, P.: The geometry of meaning: Semantics based on conceptual spaces. MIT Press (2014)
6. Gatt, A., Reiter, E.: Simplenlg: A realisation engine for practical applications. In: Proceedings of the 12th European Workshop on Natural Language Generation. pp. 90–93. Association for Computational Linguistics (2009)
7. Ghignone, L., Lieto, A., Radicioni, D.P.: Typicality-based inference by plugging conceptual spaces into ontologies. AIC@ AI* IA 1100, 68–79 (2013)

8. Harabagiu, S., Moldovan, D.: Question answering. In: The Oxford Handbook of Computational Linguistics (2003)
9. Havasi, C., Speer, R., Alonso, J.: ConceptNet: A lexical resource for common sense knowledge. Recent advances in natural language processing V: selected papers from RANLP 309, 269–280 (2007)
10. Hovy, E.: Text summarization. In: The Oxford Handbook of Computational Linguistics 2nd edition (2003)
11. Jimenez, S., Becerra, C., Gelbukh, A., Bátiz, A.J.D., Mendizábal, A.: Softcardinality-core: Improving text overlap with distributional measures for semantic textual similarity. In: Proceedings of *SEM 2013. vol. 1, pp. 194–201 (2013)
12. Lieto, A., Mensa, E., Radicioni, D.P.: A Resource-Driven Approach for Anchoring Linguistic Resources to Conceptual Spaces. In: Procs of the XV International Conference of the Italian Association for Artificial Intelligence. pp. 435–449. LNAI, Springer (2016)
13. Lieto, A., Minieri, A., Piana, A., Radicioni, D.P.: A knowledge-based system for prototypical reasoning. Connection Science 27(2), 137–152 (2015)
14. Lieto, A., Radicioni, D.P., Rho, V.: Dual PECCS: A Cognitive System for Conceptual Representation and Categorization. Journal of Experimental & Theoretical Artificial Intelligence 29(2), 433–452 (2017)
15. Lombardo, V., Piana, F., Mimmo, D., Mensa, E., Radicioni, D.P.: Semantic Models for the Geological Mapping Process, pp. 295–306. Springer International Publishing, Cham (2017)
16. Marujo, L., Ribeiro, R., de Matos, D.M., Neto, J.P., Gershman, A., Carbonell, J.: Key phrase extraction of lightly filtered broadcast news. In: International Conference on Text, Speech and Dialogue. pp. 290–297. Springer (2012)
17. Mensa, E., Radicioni, D.P., Lieto, A.: MERALI at SemEval-2017 Task 2 Subtask 1: a Cognitively Inspired approach. In: Procs. of SemEval-2017. pp. 236–240. ACL (2017)
18. Mensa, E., Radicioni, D.P., Lieto, A.: TTCS$^{\mathcal{E}}$: a Vectorial Resource for Computing Conceptual Similarity. In: EACL 2017 Workshop on Sense, Concept and Entity Representations and their Applications. pp. 96–101. ACL (2017)
19. Miller, G.A.: Wordnet: a lexical database for english. Communications of the ACM 38(11), 39–41 (1995)
20. Minsky, M.: A Framework for Representing Knowledge. In: Winston, P. (ed.) The Psychology of Computer Vision, pp. 211–277. McGraw-Hill, New York (1975)
21. Moulin, B., Irandoust, H., Bélanger, M., Desbordes, G.: Explanation and argumentation capabilities: Towards the creation of more persuasive agents. Artificial Intelligence Review 17(3), 169–222 (2002)
22. Navigli, R., Ponzetto, S.P.: BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artif Intell 193, 217–250 (2012)
23. Resnick, L.B., Salmon, M., Zeitz, C.M., Wathen, S.H., Holowchak, M.: Reasoning in conversation. Cognition and instruction 11(3-4), 347–364 (1993)
24. Rosch, E.: Cognitive Representations of Semantic Categories. Journal of experimental psychology: General 104(3), 192–233 (1975)
25. Sebastiani, F.: Machine learning in automated text categorization. ACM computing surveys (CSUR) 34(1), 1–47 (2002)
26. Tversky, A.: Features of similarity. Psychological review 84(4), 327–352 (1977)