Measuring Frame Instance Relatedness

Valerio Basile
University of Turin
Italy
basile@di.unito.it

Roque Lopez Condori
Université Côte d'Azur,
Inria, CNRS, I3S, France
roque.lopez-condori@inria.fr

Elena Cabrio
Université Côte d'Azur,
Inria, CNRS, I3S, France
elena.cabrio@unice.fr

Abstract

Frame semantics is a well-established framework to represent the meaning of natural language in computational terms. In this work, we aim to propose a quantitative measure of relatedness between pairs of frame instances. We test our method on a dataset of sentence pairs, highlighting the correlation between our metric and human judgments of semantic similarity. Furthermore, we propose an application of our measure for clustering frame instances to extract prototypical knowledge from natural language.

1 Introduction

Frame Semantics has been a staple of artificial intelligence and cognitive linguistics since its first formulation in the '70s (Fillmore, 1976). In particular, frame semantics has been widely adopted as a theoretical backbone for the interpretation of natural language, in order to represent its meaning with formal structures suited for computation. In a nutshell, according to frame semantics, the meaning of a sentence can be represented as a set of situations (*frames*) and the entities involved in them (*frame elements*), each with their own role.

Several approaches have proposed in the past years to automatically interpret natural language in terms of frame semantics (Gildea and Jurafsky, 2002; Thompson et al., 2003; Erk and Padó, 2006, among others). However, the vast majority of these approaches focuses on the extraction of the structure of the frames evoked in the natural language fragment (frames and roles), while leaving the frame elements either underspecified or simply representing them as spans of the original text. In this work, we propose to fully represent the meaning of a natural language sentence with instantiated frames, where the frame elements are nodes in a knowledge graph.

Moreover, while a great deal of effort has been directed towards the extraction of frames from natural language, not many systems process frames further, to solve downstream tasks in NLP and AI — an example is Sentilo (Recupero et al., 2015), a sentiment analysis system built on top of the frame-based machine reading tool FRED by Presutti et al. (2012).

In this paper we define a quantitative measure to compute the semantic relatedness of a pair of frame instances, and apply it to the task of creating a commonsense knowledge base.

The main contributions of this paper are:

- A novel measure of relatedness between frame instances (Section 3).
- A high-quality data set of natural language sentences aligned to the frame instances evoked by them (Sections 4 and 5).
- A pilot study on the extraction of prototypical knowledge based on frame instance clustering (Section 6).

Before introducing the novel contributions, we describe related work (Section 2), while Section 7 summarizes our conclusions.

2 Related Work

The most relevant to our research is the work of Pennacchiotti and Wirth (2009), which introduces the notion of "frame relatedness" and proposes different types of measures to asses it. These measures are grouped in three categories: i) based on the hypothesis that frames are related if their lexical units are semantically related; ii) corpus-based measures, which suggest that related frames tend to occur in

*Proceedings of the 7th Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 245–254
New Orleans, June 5-6, 2018. ©2018 Association for Computational Linguistics

the same or similar contexts (e.g., measured by pointwise mutual information or distributional semantic models); iii) hierarchy-based measures, which leverage the FrameNet hierarchy, assuming that frames are likely related if they are close in the network structure of FrameNet. The results of their experimental tests show high correlation between some of these measures and a dataset of human judgments of semantic similarity.

Subsequent works have taken the measures presented by Pennacchiotti and Wirth (2009) as basis to implement more refined measures. Kim et al. (2013) proposes SynRank, a function to calculate frame relatedness which uses three measures: i) content similarity, based on the overlapping of the terms that evoke the frames, ii) context similarity, defined by neighbor frames within a window in its document, and iii) corpus-based word similarity, which uses the corpus-specific information.

Virk et al. (2016) presented a supervised approach to enrich FrameNet's relational structure with new frame-to-frame relations. To create these new relations, the authors propose to use features based on frame network structure and frame elements (role names similarity by overlap). In addition to these features, the overlap among content words (nouns, verbs, adjectives and adverbs) occurring in verbal definitions of each frame of FrameNet is also used.

More recently, Alam et al. (2017) proposed three measures to compute the semantic relatedness between two frames using the hierarchical structure of the FrameNet graph. These measures are i) path similarity, based on the shortest path between two nodes in the taxonomy, ii) Leacock-Chodorow similarity (Leacock and Chodorow, 1998), which considers the shortest path between two nodes and the depth of the taxonomy and iii) Wu-Palmer similarity (Wu and Palmer, 1994), based on the depths of two nodes in the taxonomy and their least common subsumer. In (Shah et al.) a word sense-based similarity metric is used as a proxy to frame instance relatedness in order to cluster frame instances.

Our method presupposes a formalization of the frame element structure including the entities that fill the semantic roles, akin to the work of (Scheffczyk et al., 2006), which seeks to give the slot fillers semantic type constraints by linking them to a top-level ontology.

To our knowledge, our approach is the first to address the relatedness of instantiated frames that include disambiguated concepts in their frame elements.

3 A Quantitative Measure of Frame Instance Relatedness

In the theory of frame semantics, a *frame* is a prototypical situation uniquely defined by a name, e.g., `Driving_vehicle`, an event involving a vehicle, someone who controls it, the area where the motion takes place, and so on. Frames have *frame elements*, identified by the *role* they play in the frame. Following the example above, `Driver` and `Vehicle` are some of the frame elements expected to be present in a `Driving_vehicle` situation.

Most NLP works on frame semantics are based on FrameNet (Baker et al., 1998), a lexical semantic resource which contains descriptions and annotations of frames. In FrameNet, each frame type defines its own set of frame elements and associated words (known as *lexical units*) which can evoke the frame. FrameNet also lists a set of frame-to-frame relations (e.g. `subframe_of`, `is_causative_of`) according to how they are organized with respect to each other.

We propose a method to compute a numeric score indicating the relatedness of a pair of frame instances. Formally, we define a frame instance fi as a tuple $(f_t, \{(r_1, e_1), ..., (r_n, e_n)\})$, $f_t \in T$, $r \in R, e \in E$, where T is the set of frame types, R is the set of semantic roles, and E is the vocabulary of entities that could fill any given role.

The relatedness between two frame instances fi_1 and fi_2 is computed as a linear combination of the relatedness between the two frame types and the distance between the frame elements contained in the frame instances:

$$\begin{aligned} &firel(fi_1, fi_2) = \\ &= \alpha ftrel(fi_1, fi_2) + (1-\alpha) ferel(fi_1, fi_2) \end{aligned} \quad (1)$$

The relatedness $firel(fi_1, fi_2)$ is therefore defined to be a number in the range $[0, 1]$, while the α parameter controls the extent to which the relatedness is weighted towards the frame types or the frame elements. The frame type relatedness $ftrel$ and the frame element relatedness $ferel$ can be computed in several ways, which we detail in the remainder of this section.

3.1 Implementation Details

The method to compute the relatedness of frame instances that we propose is independent from the actual vocabulary of frames, roles and concepts — although for some of the steps precise characteristics of the frame definition are needed, e.g., a set of lexical units. In practice, we use the frame type and element inventory of FrameNet 1.5, containing 1,230 frames, 11,829 lexical units and 173,018 example sentences. As concept inventory, we select BabelNet, a large scale multilingual dictionary and semantic network (Navigli and Ponzetto, 2012). Words in BabelNet belong to one or many BabelNet synsets, each synset defines a sense, thus it represents a potential semantic role filler in a frame element.

3.2 Frame Type Relatedness

Pennacchiotti and Wirth (2009) surveys a number of methods to compute a relatedness score between frames. We implemented the best performing algorithm for frame relatedness among those introduced in the aforementioned paper, namely the *co-occurrence measure* ($ftrel_{occ}$). This algorithm is based on an estimate of the point-wise mutual information (pmi) between the two frames, computed on the basis of their occurrence in an annotated corpus.

Given two frame types ft_1 and ft_2, and a corpus C, the measure is defined as:

$$ftrel_{occ}(fi_1, fi_2) = log_2 \frac{|C_{ft_1,ft_2}|}{|C_f t_1||C_f t_2|} \quad (2)$$

where C_{ft_1} and C_{ft_2} indicate the subsets of contexts in which ft_1 and ft_2 occur respectively, and C_{ft_1,ft_2} the subset of contexts where both frame types occur.

Since a large corpus of frame-annotated natural language is hard to come by and very expensive to produce, the occurrence of a frame type ft_i in a context c is defined as the occurrence of at least one of the lexical units l_{ft_i} associated to that frame type in FrameNet in that particular context:

$$C_{ft_i} = \{c \in C : \exists l_{ft_i} \in c\}$$

$$C_{ft_1,ft_2} = \{c \in C : \exists l_{ft_1} \in c \wedge \exists l_{ft_2} \in c\}$$

While the original method only considers the word part of the lexical units, we computed the occurrence counts on SEMCOR (Landes et al., 1998), a corpus of manually sense-labeled English text (words are annotated with part-of-speech tags and senses from WordNet). By using a disambiguated corpus, we are able to match the lexical units from FrameNet to the sense labels of SEMCOR, overcoming the ambiguity of polysemous words.

We also implement an alternative measure of frame type relatedness, based on distributional semantics ($ftrel_{dist}$ inspired by another of the measures in the same paper by Pennacchiotti and Wirth (2009)). We created vector representations for each frame type by merging the representations of their lexical units in a pre-trained word space model. For each frame type, we compute the average of the vectors in GloVe6B (Pennington et al., 2014), a large word embedding model of English words, corresponding to each lexical unit in the frame. The measure of distributional frame type relatedness between two frame types ft_1 and ft_2 is then given by the cosine similarity between the two respective frame vectors $\vec{ft_1}$ and $\vec{ft_2}$:

$$ftrel_{dist}(fi_1, fi_2) = \frac{\vec{ft_1} \cdot \vec{ft_2}}{||\vec{ft_1}||||\vec{ft_2}||} \quad (3)$$

3.3 Frame Elements Relatedness

The second half of equation 1 corresponds to the relatedness measured between two sets of frame elements, therefore an aggregation step is needed. For each concept corresponding to the frame elements $fe_i \in fi_1$, we compute all the similarity scores with respect to the concepts corresponding to the frame elements $fe_j \in fi_2$, and select the best match. The aggregation by maximum is an approximation of the best match algorithm on bipartite graphs, that is, the measure gives more weight to the most similar pairs of frame elements rather than averaging the similarities of all the possible combinations. The resulting similarities are averaged over all the frame elements. Since this process is asymmetrical, we compute it in both directions and take the average of the results:

$$
\begin{aligned}
&ferel(fi_1, fi_2) = \\
&= \frac{1}{2}\Big(\frac{1}{|fi_1|} \sum_{fe_i \in fi_1} \max_{fe_j \in fi_2} csim(fe_i, fe_j) + \\
&+ \frac{1}{|fi_2|} \sum_{fe_i \in fi_2} \max_{fe_j \in fi_1} csim(fe_i, fe_j)\Big) \quad (4)
\end{aligned}
$$

The function $csim(fe_i, fe_j)$ between concepts is again computed as cosine similarity between vector representations. In this case we leverage the semantic resource NASARI (Camacho-Collados et al., 2016), a concept space model built on top of the BabelNet semantic network. Each vector in NASARI represents a BabelNet synset in a dense 300-dimensional space. The reason to use a different vector space model than the one used for $ftrel_{dist}$ is that NASARI provides representations of disambiguated concepts, which we have from KNEWS, while GloVe6B is a word-based model and the lexical units are not disambiguated.

Note that in equation 4 the semantic roles of the elements are ignored in the computation of the relatedness between frame elements. We therefore extend the definition of frame element relatedness by adding the extra parameter $roles$, acting as a filter: when activated, it sets the relatedness score of a pair of frame elements to zero if they do not share the same role in the frame instance.

4 Evaluation by Text Similarity

To our knowledge, there is no manually annotated dataset of frame instances and their relatedness. In order to circumvent this shortcoming, we propose an indirect methodology for the evaluation of the frame instance relatedness measures we introduced in Section 3. The key idea of our evaluation approach is to measure the relatedness of frame instances extracted from pairs of short texts, for which a gold standard pairwise similarity score is given.

We parse the text with a knowledge extraction system to extract all the frame instances. We then measure the semantic relatedness of the extracted frame instances and compare the outcome with a judgment of pairwise semantic similarity given on the original sentences. The aim of this experiment is to show that our measure of frame instance relatedness correlates with the semantic relatedness of the text that evokes the frame. In other words, we use textual similarity as a proxy for human judgment of relatedness between frame instances.

4.1 Data

The dataset we selected to carry out this experiment is provided by the shared task on Semantic Text Similarity (STS) held at SemEval 2017 (task 1, track 5 English-English) (Cer et al., 2017). The set is composed of 250 pairs of short English sentences, manually annotated with a numerical score from 1 to 5 indicating their degree of semantic relatedness. Examples of sentence pairs from the gold standard set, along with their human judgments of semantic similarity, are shown in Table 1.

Table 1: Examples of the sentence pairs in the SemEval 2017 STS dataset, with numbers indicating their semantic similarity on a scale from 1 to 5.

Sim.	Sentence pair
4.0	There are dogs in the forest. The dogs are alone in the forest.
3.4	The boy is raising his hand. The man is raising his hand.
1.0	A woman supervisor is instructing the male workers. A woman is working as a nurse.
0.2	The woman is kneeling next to a cat. A girl is standing next to a man.

4.2 Knowledge Extraction

To compute the relatedness score of pairs of frame instances, we need to extract them from the natural language text. For this purpose, we use KNEWS (Knowledge Extraction With Semantics), a fully automated pipeline of NLP tools for machine reading (Basile et al., 2016). The input of KNEWS is an arbitrary English text, and its output is a set of RDF triples encoding the frames extracted from the text by Semafor (Das et al., 2014). KNEWS integrates the Word Sense Disambiguation tool Babelfy (Moro et al., 2014) to extract concept and entities from the input sentences, and maps them to the frame roles provided by Semafor, creating frame instances where the frame types are from FrameNet 1.5 and the frame roles are filled with concepts from BabelNet. An example of the extraction of frame instances from natural language performed by KNEWS is shown in Figure 1. In the example, three frame instances are extracted from the sentence "two men sit on a bench", with frame types `People`, `Cardinal_numbers` and `Being_located`. The frame elements are completed with BabelNet synset identifiers, e.g., the `Theme` of `Being_located` is `bn:00001533n` (man, adult male, male: *An adult person who is male (as opposed to a woman)*[1]) and the `Location` of the same frame instance is `bn:00009850n` (bench: *A long seat for more than one person*[2]).

We ran KNEWS on the 500 sentences from the STS dataset and extracted 1,650 frame instances of 178 different frame types. Each frame instance has on average 1.2 frame elements, for a total of 2,107 roles filled by 457 different types of concepts.

4.3 Frame-based Sentence Similarity

Our aim in this experiment is to assess the relatedness of *sentences* by measuring the relatedness of their corresponding frame instances. Since we have defined (in Section 3) a method to compute the relatedness of *frame instances*, an extra step of aggregation is needed in order to reconcile the measurement for the evaluation. We define the similarity $ssim(s_1, s_2)$ between two sentences $s_1 = \{fi_1^1, ..., fi_n^1\}$ and $s_2 = \{fi_1^2, ..., fi_m^2\}$ as follows:

$$
\begin{aligned}
& ssim(s_1, s_2) = \\
& = \frac{1}{2}\Big(\frac{1}{|s_1|}\sum_{fi_i^1 \in s_1} \max_{fi_j^2 \in s_2} firel(fi_i^1, fi_j^2) + \\
& + \frac{1}{|s_2|}\sum_{fi_i^2 \in s_2} \max_{fi_j^1 \in s_1} firel(fi_i^1, fi_j^2)\Big) \quad (5)
\end{aligned}
$$

Table 2: Pearson correlation between sentence pair similarity scores predicted by frame instance relatedness and the SemEval STS reference set.

	without role filter		with role filter	
ftrel:	**occ**	**dist**	**occ**	**dist**
alpha				
1.0	0.526	0.455	0.526	0.455
0.9	0.529	0.465	0.536	0.477
0.8	0.529	0.471	0.544	0.495
0.7	0.525	0.473	0.550	0.510
0.6	0.517	0.471	0.555	0.522
0.5	0.503	0.463	0.558	0.531
0.4	0.484	0.451	0.558	0.538
0.3	0.461	0.436	0.557	0.542
0.2	0.436	0.418	0.554	0.544
0.1	0.410	0.400	0.550	0.545
0.0	0.381	0.381	0.543	0.543

We tested the effect of the α parameter, the frame type relatedness measures $ftrel_{occ}$ and $ftrel_{dist}$, and the filter on semantic roles to investigate their impact on the quality of the relatedness measurement. The result is given in Table 2 in terms of Pearson correlation between the gold standard relatedness scores and the relatedness scores predicted by our method.

Overall, the occ measure of frame type relatedness produces better results than $dist$. We find that both halves of equation 1 contribute to the final result, with a sweet stop around $\alpha = 0.4$ that achieves the best performance on this benchmark with $ftrel = occ$ and the filter on the semantic roles. Indeed, enforcing the matching constraint on the semantic roles proves to be a successful strategy. The difference in terms of adherence to the text similarity scores with and without such constraint is significant and consistent across every variation of the other parameters.

4.4 Discussion

It must be stressed that the aim of the experiment presented in this section is not to achieve state of the art performance on the STS task, for which better algorithms based on word similarity and other techniques have been proposed. In fact, many tasks that rely on sentence level semantics can be solved without the need of extracting frame instances. Rather, we show that our method to compute a relatedness score between frame instances works

[1] http://babelnet.org/synset?word=bn:00001533n

[2] http://babelnet.org/synset?word=bn:00009850n

```
@prefix fbfi: <http://framebase.org/ns/fi->
@prefix fbframe: <http://framebase.org/ns/frame->
@prefix fbfe: <http://framebase.org/ns/fe->
@prefix rdfs: <http://www.w3.org/1999/02/22-rdf-syntax-ns\#>
@prefix bn: <http://babelnet.org/rdf/>

fbfi:People_01b52400 rdfs:type fbframe:People.
fbfi:People_01b52400 fbfe:Person bn:00001533n.
fbfi:Cardinal_numbers_3faa6c9c rdfs:type fbframe:Cardinal_numbers.
fbfi:Cardinal_numbers_3faa6c9c fbfe:Entity bn:00001533n.
fbfi:Being_located_079aed4d rdfs:type fbframe:Being_located.
fbfi:Being_located_079aed4d fbfe:Theme bn:00001533n.
fbfi:Being_located_079aed4d fbfe:Location bn:00009850n.
```

Figure 1: Frame instances extracted by KNEWS from the sentence "two men sit on a bench".

in practice, despite the inevitable shortcomings of the frame extraction process, i.e., wrong and/or missing classifications of frames, roles and concepts. The STS dataset has a strong bias towards people-centric frames. In fact, the most frequent frame type in our collection is People (345 occurrences in 1,650 frame instances), and the most frequent concept is `bn:00001533n` (*man, adult male, male*, 226 occurrences in 2,107 frame elements).

5 Evaluation on Gold Standard Frame Instances

The evaluation conducted in the first experiment has the advantage of being fully automated. However, measuring frame instance relatedness indirectly through text similarity entails that two distinct effects are measured at once: 1) the relatedness of the frame instances extracted from the text, and 2) the accuracy of the frame instance extraction process. In this section we propose a revised methodology for the evaluation of the frame instance relatedness measure that focuses only on measuring the effect (1), canceling the interference of (2). In short, we manually correct the frame instances extracted with KNEWS from the STS sentence pairs and re-run the evaluation process as described in Section 4. As by-products, we create a gold standard dataset of frame instances aligned with the text that evokes them[3], and we provide an evaluation of the performance of the KNEWS knowledge extraction system.

5.1 Manual correction

We corrected each frame instance individually. For the frame types, they were either confirmed or marked as wrong. In the latter case, the frame instance is discarded from the data set without further process. This was also the procedure applied when an entity was not filling any role for a particular frame instance, due to a parsing mistake. If the frame type was confirmed by the annotator, then the role and sense labels were checked and possibly corrected by replacing them with the correct ones from FrameNet and BabelNet respectively.

We split the STS dataset (250 sentence pairs) in three parts and assigned each of them to an annotator. A subset of 37 frame instances extracted from 10 sentences was annotated by all three annotators in order to compute a measure of inter-coder reliability, resulting in a Fleiss' Kappa of 0.81 on the annotation of frame types, 0.76 for roles, and 0.90 for concepts. Note that the annotation of roles and concepts is only considered when frame types are not discarded by the annotators as wrong.

Once the annotation was finished, we compared the obtained dataset with the one we produced with KNEWS (Section 4.2). The accuracy at the frame instance level (rate of frame instances that were not corrected at all) is 77.1%. More in detail, 79.5% of the frame types were found correct. Among the frame instances with correct frame types, 95.9% of the roles and 82.5% of the concepts were correct. During the manual inspection, we confirmed that Semafor (like most semantic parsers) is

[3] We will release the dataset after the review period.

biased towards the most dominant frame for ambiguous forms. The final gold standard set comprises 1,261 frame instances and 1,579 frame elements.

5.2 Text Similarity Experiment with Gold Standard Frame Instances

We repeated the experiment in 4.3, this time computing the pair-wise frame relatedness on the manually corrected frame instances. To provide a fair comparison, we removed the frame instances from the original set corresponding to the frame instances removed during the manual correction. We used the filter on semantic roles described in 4.3 and $ftrel_{occ}$ (the performance patterns we observed were the same as in the original experiment). The results of the experiment are shown in Figure 2. The overall performance is slightly lower than the previous experiment. This can be explained by observing that in this version of the experiment we are using less data, although of higher quality. Due to the structure of the correlation-based evaluation, incorrect frame instances extracted from a pair of sentences contribute to their relatedness score more than missing some frame instances. Also, the dominance bias could play a role, in that we mostly discarded low-frequency frames, for which the relatedness metric we defined could perform less than optimally. An in-depth analysis of this phenomenon (i.e., how does lexical ambiguity interplay with the variance in relatedness scores?) is left for future work.

6 Clustering Frame Instances to Extract Prototypical Knowledge

In the previous sections, we proved that our method for computing a relatedness score between two frame instances correlates well with human judgments of semantic similarity based on the natural language expression of such instances. What we presented is a kind of intrinsic evaluation, which, while helpful in assessing the quality of the solution, does not provide an insight into the motivation to implement a measure of frame instance relatedness, and what open problems could benefit from our approach down the line. To fill this gap, we propose a pilot study on the application of the method introduced in this paper to a downstream task, namely the extraction of common sense knowledge from text, in line with the prototypical knowledge building method in (Shah et al.).

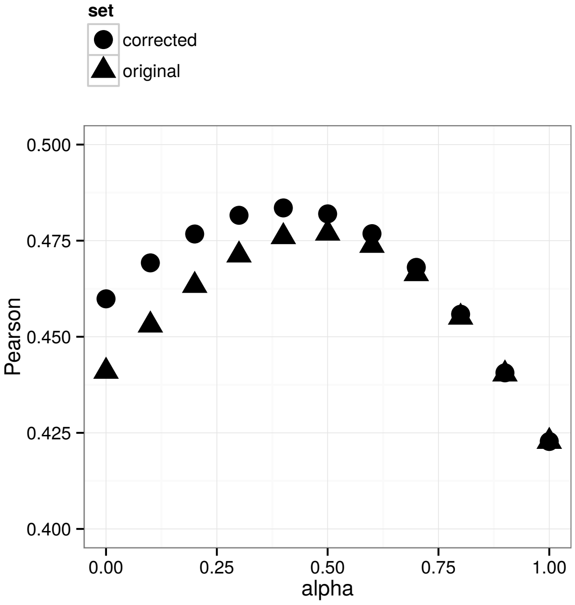

Figure 2: Pearson correlation between sentence pair similarity scores predicted by frame instance relatedness on corrected frame instances and the SemEval STS reference set.

We start by observing that defining a quantitative distance metric between homogeneous instances allows us to apply a clustering algorithm. The result of such clustering is a partition of the original set into subsets that can be either overlapping (*soft clustering*) or non-overlapping (*hard clustering*). Moreover, clusters have a definite shape, with one of the elements being the most central one (called the *clustroid*), and the others being more or less far from the center. We perform a hard clustering of the frame instances collected from the STS dataset and used for the experiment in Section 4.3, and formulate three hypotheses: i) elements close to the center of their respective clusters are the best candidates to represent prototypical frame instances; ii) elements near the border of their respective clusters are less likely to represent prototypical frame instances, and therefore can be filtered out; iii) the size of each cluster influences the prototypicality degree of the elements in its central region, with larger clusters containing more prototypical frame instances near its center.

To cluster the frame instances, we follow

Table 3: Random sample of frame instances extracted from the STS dataset.

Cluster size	5
Frame type	`Noise_makers` (the Noise_maker is an artifact used to produce sound, especially for musical effect)
Role	`Noise_maker` (this FE identifies the entity or substance that is designed to produce sound)
Concept	`Guitar` (a stringed instrument usually having six strings; played by strumming or plucking)
Cluster size	40
Frame type	`Substance` (this frame concerns internally undifferentiated Substances)
Role	`Substance` (the undifferentiated entity which is presented as having a permanent existence)
Concept	`Sand` (a loose material consisting of grains of rock or coral)
Cluster size	3
Frame type	`Part_inner_outer` (This frame concerns Parts of objects that are defined relative to the center or edge of the object
Role	`Part`
Concept	`Center` (an area that is approximately central within some larger region)
Role	`Whole` (an undivided entity having all its Parts)
Concept	`Pond` (a small lake)

Table 4: Clustroids of randomly selected clusters from the STS dataset.

Cluster size	8
Frame type	`Vehicle` (the frame concerns the vehicles that human beings use for the purpose of transportation)
Role	`Vehicle` (is the transportation device that the human beings use to travel)
Concept	`Boat` (a small vessel for travel on water)
Cluster size	5
Frame type	`Biological_area` (this frame contains words that denote large ecological areas as well as smaller locations characterized by the type of life present)
Role	`Locale` (this FE identifies a stable bounded area)
Concept	`Forest` (the trees and other plants in a large densely wooded area)
Cluster size	35
Frame type	`Roadways` (This frame involves stable Roadways which connect two stable Endpoints, the Source and the Goal)
Role	`Roadway` (the Roadway is the roadway that connects locations)
Concept	`Road` (a way or means to achieve something)

Table 5: Clustroids of the three largest clusters in the dataset.

Cluster size	418
Frame type	`People` (this frame contains general words for Individuals, i.e. humans)
Role	`Person` (the Person is the human being)
Concept	`Man` (an adult person who is male -as opposed to a woman-)
Cluster size	51
Frame type	`Clothing` (this frame refers to clothing and its characteristics, including anything that people conventionally wear)
Role	`Garment` (this FE identifes the clothing worn)
Concept	`Shirt` (a garment worn on the upper half of the body)
Cluster size	50
Frame type	`Kinship` (this frame contains words that denote kinship relations)
Role	`Alter` (the person who fills the role named by the Kinship term with respect to the Ego)
Concept	`Child` (a young person of either sex)

the hierarchical clustering approach, because the number of clusters is not necessary to be known a priori. In particular, we used the version implemented in the SciPy library[4]. We tested different linkage methods for hierarchical clustering (single, complete, average, weighted, centroid, median and ward), observing comparable results in terms of number of clusters and their size distribution. We perform the clustering with average linkage and the best performing parameters of the frame relatedness

[4] https://www.scipy.org/

measure (the distance metric for the clustering) according to the experiments in Section 4.

While giving an objective assessment about the prototypicality of a frame instance is somewhat hard, we observe different behavior in line with our hypothesis. The examples reported in Table 3 include quite arbitrary, albeit correct, frame instances. On the other hand, the examples in Table 5 are indeed highly prototypical, e.g., a shirt is a prototypical piece of clothing, while the examples in Table 4 can be placed somewhere in the middle of the prototypicality scale.

7 Conclusion and Future Work

We presented a novel method to compute a quantitative relatedness measure between frame instances, that takes into account the type of the frames, the semantic role of the frame elements, and the entities involved in the frame instances. Based on a test conducted on a gold standard set of sentence pairs, the measure we defined correlates positively with human judgments of semantic similarity. We further apply the relatedness measure to the task of extracting prototypical knowledge from natural language.

One clear bottleneck of our experimental setup is given by the automatic parsing, that does not always reach optimal performances. We believe that a stable measure of relatedness between frame instances will in fact boost the performance of a disambiguation system, acting as a coherence measure for an all-word disambiguation approach. We intend to test such strategy in future work.

The experiment on frame instance clustering for prototypical knowledge extraction presented in Section 6 showed promising results. In future work, we plan to conduct a large-scale experiment following the same principles including an extensive systematic evaluation of the quality of the resulting dataset.

References

Mehwish Alam, Diego Reforgiato Recupero, Misael Mongiovì, Aldo Gangemi, and Petar Ristoski. 2017. Event-based Knowledge Reconciliation using Frame Embeddings and Frame Similarity. *Knowledge-Based Systems*, 135:192–203.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet Project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1*, ACL '98, pages 86–90, Stroudsburg, PA, USA. Association for Computational Linguistics.

Valerio Basile, Elena Cabrio, and Claudia Schon. 2016. KNEWS: Using Logical and Lexical Semantics to Extract Knowledge from Natural Language. In *Proceedings of the European Conference on Artificial Intelligence (ECAI) 2016*.

José Camacho-Collados, Mohammad Taher Pilehvar, and Roberto Navigli. 2016. Nasari: Integrating explicit knowledge and corpus statistics for a multilingual representation of concepts and entities. *Artificial Intelligence*, 240:36–64.

Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. 2017. Semeval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, Vancouver, Canada.

Dipanjan Das, Desai Chen, André F. T. Martins, Nathan Schneider, and Noah A. Smith. 2014. Frame-semantic parsing. *Computational Linguistics*, 40:1:9–56.

Katrin Erk and Sebastian Padó. 2006. SHALMANESER - A Toolchain For Shallow Semantic Parsing. In *Proceedings of LREC 2006*.

Charles J. Fillmore. 1976. Frame semantics and the nature of language. *Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech*, 280(1):20–32.

Daniel Gildea and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Comput. Linguist.*, 28(3):245–288.

H. Kim, X. Ren, Y. Sun, C. Wang, and J. Han. 2013. Semantic Frame-Based Document Representation for Comparable Corpora. In *Proceedings of the 13th International Conference on Data Mining*, pages 350–359.

Shari Landes, Claudia Leacock, and Randee I Tengi. 1998. Building semantic concordances. *WordNet: An electronic lexical database*, 199(216):199–216.

C. Leacock and M. Chodorow. 1998. Combining local context and wordnet similarity for word sense identification. In *MIT Press*, pages 265–283, Cambridge, Massachusetts.

Andrea Moro, Alessandro Raganato, and Roberto Navigli. 2014. Entity Linking meets Word Sense Disambiguation: a Unified Approach. *Transactions of the Association for Computational Linguistics (TACL)*, 2:231–244.

Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Marco Pennacchiotti and Michael Wirth. 2009. Measuring Frame Relatedness. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, pages 657–665, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Valentina Presutti, Francesco Draicchio, and Aldo Gangemi. 2012. Knowledge extraction based on discourse representation theory and linguistic frames. In *Knowledge Engineering and Knowledge Management*, pages 114–129, Berlin, Heidelberg. Springer Berlin Heidelberg.

Diego Reforgiato Recupero, Valentina Presutti, Sergio Consoli, Aldo Gangemi, and Andrea Giovanni Nuzzolese. 2015. Sentilo: Frame-based sentiment analysis. *Cognitive Computation*, 7(2):211–225.

Jan Scheffczyk, Adam Pease, and Michael Ellsworth. 2006. Linking framenet to the suggested upper merged ontology. In *Proceedings of the 2006 Conference on Formal Ontology in Information Systems: Proceedings of the Fourth International Conference (FOIS 2006)*, pages 289–300, Amsterdam, The Netherlands, The Netherlands. IOS Press.

Avijit Shah, Valerio Basile, Elena Cabrio, and Sowmya Kamath S. Frame instance extraction and clustering for default knowledge building. pages 1–10.

Cynthia A Thompson, Roger Levy, and Christopher D Manning. 2003. A generative model for semantic role labeling. In *European Conference on Machine Learning*, pages 397–408. Springer.

Shafqat Mumtaz Virk, Philippe Muller, and Juliette Conrath. 2016. A Supervised Approach for Enriching the Relational Structure of Frame Semantics in FrameNet. In *Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3542–3552, Osaka, Japan. The COLING 2016 Organizing Committee.

Zhibiao Wu and Martha Palmer. 1994. Verbs semantics and lexical selection. In *Proceedings of the 32Nd Annual Meeting on Association for Computational Linguistics*, ACL '94, pages 133–138, Stroudsburg, PA, USA. Association for Computational Linguistics.