# Tackling Frame-Semantic Role Labeling through Analogies

**Anonymous ACL submission**

## Abstract

Analogy making has been claimed to be at the core of cognition and be intimately related with abstraction and categorization. Despite its importance, analogies have only been scantily studied in Natural Language Processing (NLP) with most work being limited to word analogies. Most extant approaches view analogy making as the identification of the fact that pairs $(A, B)$ and $(C, D)$ share the same latent relation without necessarily naming that relation. In this paper we adapt this framework in the context of frame semantics, focusing on the problem of semantic role labeling. For a given target sentence and a predicate we are able to identify all semantic roles, casting the problem as an analogy with a previously annotated sentence of the same frame, but not necessarily of the same predicate. We show that careful selection of the source sentence has the potential to surpass state of the art results while at the same time using a computationally lean model.

## 1 Introduction

It has been claimed that analogies play a crucial role for human cognition (*cf* Hofstadter, 2001; Hofstadter and Sander, 2013, *inter alia*). Analogies can be viewed as the abstraction mechanism that identifies common salient features between two objects. They are usually represented as a relation between two pairs $A : B :: C : D$ (usually read *A is to B as C is to D*). A common assumption that existing approaches on analogies in NLP make is that pairs $(A, B)$ and $(C, D)$ share the same latent relation. The underlying elements $A, B, C, D$ of the analogy can be words (Turney, 2008), their vectorial representations (Mikolov et al., 2013b,a) or larger chunks of text such as sentences (Zhu and de Melo, 2020; Afantenos et al., 2021, 2022).

In this paper our goal is to introduce a new framework leveraging analogies in order to identify latent relations between elements in a sentence. We use FrameNet 1.7 (FN1.7, Baker et al., 1998) as our testbed which essentially provides a lexicon of *semantic frames* as well as a set of annotated sentences with frames from this resource. Each frame represents an event or state which is *triggered* by a specific word or expression in the sentence. Each frame contains a set of *semantic roles* (SR, or frame elements) which can be mandatory (core semantic roles) or not (peripheral or extra-thematic semantic roles).

Frame-semantic parsing is a series of three sequential tasks: identification of all words or expressions that trigger a frame (also known as *predicates*, identification of the frame or frames that are triggered, and finally for each frame $f$ and the set $\mathcal{R}_f$ all possible semantic roles as defined in that frame associate a text span of the sentence with each role, if such an association exists. The last task is known as *argument identification and classification* or Frame-Semantic Role Labeling (FSRL). Current state of the art approaches use sophisticated encoding (graph neural representations, Lin et al., 2021) or decoding mechanisms (semi-Markov CRFs, Swayamdipta et al., 2017) in order to perform full semantic parsing.

In this paper we concentrate solely the task of FSRL and we cast it as an analogy solving problem. More specifically, given a source sentence and a pair $(A, B)$ representing the predicate triggering frame $f$ and argument evoking a specific semantic role in that sentence, as well as a target sentence and its predicate $C$ that triggers the same role $f$ we seek to identify $D$ in the target sentence such that $B$ and $D$ evoke the same semantic role. We show that by using analogies we can obtain results that outperform state of the art under some circumstances without having recourse to sophisticated encoding or decoding mechanisms.

Our main contributions in this paper are:

- a new task for studying analogical proportions between groups of words from different sentences, at the frame semantics level, moving towards better understating of analogies between sentences;

- a dataset of semantic analogies that complements the traditional datasets of analogies between words on factual and lexical semantics (Drozd et al., 2016);
- we show that a simple model based on semantic embeddings is enough to solve analogies with a high performance;
- we demonstrate how a computationally inexpensive model can exploit analogies in order to achieve results that can outperform the state of the art to tackling FSLR.

## 2 Related work

Turney (2008) introduced *Latent Relational Analysis (LRA)* in order to identify such analogies which he tested in 20 scientific and metaphorical examples . Later Mikolov et al. (2013b,a) used analogies in order to evaluate the quality of the word embeddings produced with *word2vec*. Since then word analogies have been widely used to evaluate the intrinsic qualities of word embeddings, although it has been shown that this is not sufficient since most models appear to take shortcuts instead of learning abstraction and analogical mapping. In particular, Gladkova et al. (2016) that not-well balanced datasets, such as the Google analogy test set (Mikolov et al., 2013b,a), do not permit us to safely conclude that underlying embeddings combined with the vector offset approach are able to capture analogies. They introduce the Bigger Analogy Test Set (BATS) showing that derivational and lexicographic relations remain a challenge. Rogers et al. (2017) show that the vector offset approach as well as 3CosAdd (Levy and Goldberg, 2014) suffer from dependence on vector similarity arguing against the use of such datasets in order to evaluate the intrinsic values of word embeddings.

Sultan and Shahaf (2022) adapt the framework of the *Structure Mapping Theory* (Gentner, 1983) on procedural texts, extracting entities and their relationships finding a mapping between two different domains based on relational similarity. Relationships are sets of ordered verbs and between entities, which are extracted based on question/answer pairs. Similarity measures the fact that two sets share more relations. Mappings are identified heuristically based on the cosine similarity of the Bert vectors representing the questions that provided the entities. Their approach successfully extracts mappings in two different datasets.

To the best of our knowledge analogies have not been used in the context of FSRL Swayamdipta et al. (2017) present a softmax-margin semi-Markov model. The authors use a bidirectional-RNN with a semi-Markov CRF without initially using any syntactic features. They then use multi-task learning and syntactic scaffolding obtaining state of the art results at the time of publication. More recently, Lin et al. (2021) use Graph Neural Networks based on Bert embeddings and BiHLSTMs (Srivastava et al., 2015) for the full frame semantics task obtaining also state of the art results.

## 3 Methodology

In this section, we formulate our analogy solving problem on the identification of *semantic roles (SRs)* and describe the model used to tackle it. Contextual information is necessary to understand the semantic role of a group of words, as it is defined in relation to a semantic frame. Accordingly, we decided to focus on contextualized word embedding models, in particular mBert (Devlin et al., 2019) as we intend to expand our application to other languages in further work.

### 3.1 Problem formulation

As mentioned in the introduction, we focus solely on the task of FSRL. Given a target sentence and a predicate of that sentence as well as the frame $f$ that is evoked from that predicate we seek to identify all spans of text in the sentence that are associated with role $r \in \mathcal{R}_f$. In order to do so, we select another source sentence with a predicate (not necessarily the same) that triggers the same frame and has already been annotated with all its semantic roles and cast the problem as an analogy solving one. More specifically, given a source sentence $s = \{w_1^s, \ldots, w_n^s\}$ and a distinct target sentence $t = \{w_1^t, \ldots, w_m^t\}$ each represented by their sequence of tokens, we will consider three substrings of consecutive tokens in $s$ and $t$ respectively

$$A = \{w_{i_A}^s, \ldots, w_{i_A+|A|-1}^s\},$$

$$B = \{w_{i_B}^s, \ldots, w_{i_B+|B|-1}^s\},$$

$$C = \{w_{i_C}^t, \ldots, w_{i_C+|C|-1}^t\},$$

with $i_A, i_B, i_C$ representing the index of the starting word position for $A, B, C$ respectively. $A$ and $B$ belong to $s$ while $C$ belongs to $t$. $A$ and $C$ represent the predicates that trigger the same frame $f$

in $s$ and $t$. We seek to identify $D = \{w_i^t, \dots, w_j^t\}$ with $i, j \in [1, m]$ and $i \leq j$ such that $B, D$ activate the same semantic role $r \in \mathcal{R}_f$. In other words, $A, B, C, D$ form a valid analogy $A : B :: C : D$, and we are looking to solve the analogical equation $A : B :: C : x$ (Prade and Richard, 2021).

### 3.2 Model formalization

We define two probability distributions $p_b, p_e$ over the tokens of $t$, respectively the likelihood of a token being the first token of the answer (the **b**eginning) and the last token of the answer (the **e**nd). The two probability distributions are conditioned by $s, t$ as well as by the analogy $A : B :: C : x$ we want to solve. Then, the analogy solving problem can be formulated as follows, with $i \leq j$:

$$\underset{i,j\ \in[0,m]}{\text{argmax}}\{p_b(w_i^t|s,t,A,B,C) + p_e(w_j^t|s,t,A,B,C)\}. \tag{1}$$

Conditional probabilities for each word being the start or end of fourth element of an analogy given the two sentences and the first three elements of the analogy, are obtained using the pretrained transformer architecture mBert (Devlin et al., 2019, cf. also Appendix D.1) fine tuned using the proposed *extractive question answering (Ex-QA)* model for solving SQuAD[1] (Rajpurkar et al., 2016).

For each word[2] $w_i \in t$ we obtain contextual embeddings

$$\mathbf{w}_i = mBert(w_i|s,t,A,B,C)$$

which we then feed to two single layer neural networks learning whether a token constitutes the beginning or end of a segment which is a solution to an analogy. More specifically, we estimate $\mathbf{z_b}(i) = \mathbf{W}_b^T\mathbf{w}_i + \mathbf{b_b}$ and $\mathbf{z_e}(i) = \mathbf{W}_e^T\mathbf{w}_i + \mathbf{b_e}$ where $\mathbf{W}_b, \mathbf{W}_e, \mathbf{b}_b, \mathbf{b}_e$ are learned matrices. Conditional probabilities are obtained for each token given the context using a $softmax$ function:

$$p_b(w_i^t|s,t,A,B,C) = \frac{e^{\mathbf{z}_b(i)}}{\sum_{j\in t} e^{\mathbf{z}_b(j)}},$$

$$p_e(w_i^t|s,t,A,B,C) = \frac{e^{\mathbf{z}_e(i)}}{\sum_{j\in t} e^{\mathbf{z}_e(j)}}.$$

During decoding we require $i \leq j$ but no further constraints are imposed.

Notice that in Eq. 1 it is possible to have $i, j = 0$. Inspired by (Devlin et al., 2019) we consider a special token $w_0^t$ which helps us handle instances in which no analogy exists. This is the case when the optimal solution for Eq. 1 yields $i = j = 0$, denoting a *negative instance*, detailed in §4.1. Otherwise, we consider only $0 < i \leq j$ during decoding.

## 4 Experiments

We perform two kinds of experiments. In §4.1 we analyze the performance of the analogy solving model we developed in different settings, and explore the sensitivity of our model to perturbations on key aspects of the approach. These experiments allow us to confirm the soundness of the approach with regards to the analogy solving process. Then, in §4.2, we apply our analogy solving model to FSRL, and show the potential of our approach to outperform state of the art model, with the added benefit of a relative simplicity of our approach compared to the complex architecture of the state of the art models.

### 4.1 Analogy solving performance

Following the formulation introduced in §3.1, we train an analogy solving model on the training data described hereafter. We determine the limitations of our model with regards to the analogical setting, and conclusions drawn here can be transferred to the FSRL setting. Indeed, the analogies used for FSRL in §4.2 are a special case of the ones used here by considering only $A, C$ as frame predicates.

**Dataset.** To train our model and explore its analogy solving performance, we use a dataset built upon FN1.7 containing analogies involving instances of core SRs and, in some cases, the frame predicate. The dataset is detailed in Appendix C.1, including key aspect of dataset construction.

As mentioned in §3.2, it is possible that some SRs of a given frame are not instantiated in a given sentence. To account for this, we consider positive instances of analogical equation that can be solved because $r_D$ is instantiated in $t$ as $D$, and negative instances where the $r_D$ is not instantiated in $t$ and the equation cannot be solved.

**Training hyperparameters.** The model is trained for at most 1 epoch, and early stopping

[1] https://huggingface.com

[2] The Bert architecture considers *tokens* which are different from *words* in the linguistic meaning, as for instance a word may be split in multiple tokens and tokens can be punctuation marks. Still, we prefer to use word in this paper to facilitate reading.

is decided on the development set, using an approximation of the Word Error Rate (WER) using the token positions, coined Token Position Error Rate (TPER) and detailed in Appendix C.2. Batch size is automatically found by the Huggingface library to maximize GPU usage.

**Evaluation method.** For all instance classes, we report the accuracy of the model, which is the percentage of instances where the model returns the expected output (the *gold SR* for positive instances, or the $w_0^{CD}$ token for negative instances). If the model does not return the expected output, we speak of model *failure* and consider 3 possibilities: "*wrong SR*" if the model returns a instance of an SR that is different from the gold SR; "*SR not found*" if the model outputs $w_0^t$ even if the analogy could be solved (*i.e.*, positive instances); any other case corresponds to outputs that do not exactly match an SR nor the $w_0^t$ span, that we call "*not an SR*". Note that *wrong SR*, *SR not found*, and *not an SR* cover all the possible cases of model failure, so *accuracy + wrong SR + SR not found + not an SR* = 100%.

| Instances | Accuracy | Wrong SR | Not an SR | SR not found |
|---|---|---|---|---|
| *Analogical model (using A,B,C)* | | | | |
| Positive | 72.31% | 0.28% | 11.24% | 16.17% |
| Negative | 72.09% | 0.52% | 27.38% | – |
| All | 72.21% | 0.39% | 18.43% | 8.97% |
| $r_A \neq r_C$ | *70.75%** | 0.31% | 11.59% | *17.36%** |
| *Non-analogical model (using only B)* | | | | |
| Positive | *53.46%** | 0.10% | 16.59% | *29.85%** |
| Negative | 75.32% | 0.39% | 24.30% | - |
| All | 63.19% | 0.23% | 20.02% | 16.56% |

Table 1: Analogy solving results (in % of all instances) for the analogical and non-analogical models. Instances where $r_A \neq r_C$ are not counted in the overall performance (*All*).

**Overall performance.** We report in Table 1 the performance of our model on positive instances (solvable analogies) and negative instances (unsolvable analogies due to missing SR instance), as well as the average performance over those two classes of instances. *SR not found* is not given for negative instances, as it is the expected output.

Firstly, there is no significant difference between the accuracy on positive and negative instances, with a high level of performance (above 72% accuracy) in both cases. For reference, Djemaa et al. (2016) report 77% inter annotator agreement for roles of matching frames. With positive instances, the model wrongly determines that the SR is not instantiated in only about 16% of cases. However, for negative instances the model errors are almost exclusively *not an SR*. As missing annotations and errors are present in the part of FN1.7 we use, it is likely that some of our instances are solvable but the instance for $r_D$ is not labeled, so the instances are counted as negative ones. Similarly, it is possible that some measured errors are better than the recorded annotation, but checking this hypothesis requires manually checking samples for annotation errors which is beyond the scope of the article. Overall, our model has a high accuracy, despite the punitive way we determine failures: in the case of multi-token words, the models fails if a token part of a word is omitted while the other tokens of the word is correctly predicted, and conversely for tokens that are wrongly predicted.

Secondly, to obtain a deeper understanding of the errors made by the model we consider the Safe Word Error Rate (SafeWER), a slight modification of the WER to handle the empty targets we have for our negative instances. In Appendix C.3, we provide the formula of SafeWER, provide intuitions about its meaning, and report the results for our analogy solving models. When the model does not correctly predict the solution of the analogy, the SafeWER is significant, with in average 0.96 for positive instances (in average, about as many mistakes as the number of expected words) and 2.53 for negative instances (in average 2.5 words predicted when failing to identify a non solvable analogy). The SafeWER we obtain and the *not an SR* instances indicate that a significant part of our model mistakes is due to few extra tokens or forgotten tokens. In particular, it is interesting that a significant part of mistakes are *not an SR* while *wrong SR* is very rare, as it indicates that in many cases the issue comes from the identification of the boundaries of the SR and not from the identification of the fitting SR itself. Adding span identification into our model, as is done in multiple FSRL systems (Lin et al., 2021; Zheng et al., 2022), should significantly improve performance in that area.

**Impact of $A, C$.** We study the sensibility of the model to several perturbations regarding the SRs $A, C$ for two purposes: we measure the impact of $A, C$ on the performance of the model from the analogical point of view, and, by extension, the impact of errors in identifying the frame predicate on the FSRL performance. To do so, we generate analogies such that $r_A \neq r_C$ but $r_B = r_D$, which means that while the analogy is erroneous, it is still possible to solve it. Additionally, while $r_A \neq r_C$, $r_A, r_C$ are kept as instances of the same frame as $r_B, r_D$. The generation process is described in Appendix C.1.

Results are shown in the $r_A \neq r_C$ column of Table 1, with the points of interest marked with "*". First, while there is a drop in accuracy, the performance remains very high, with only a 2% decrease. Additionally, it is interesting to see that the new errors mostly belong to the "SR not found". While the difference might not be significant enough to draw conclusions, we propose the following hypothesis: by introducing a mismatch $r_A \neq r_C$ in the starting point of the relation, the model determines that there is no instance that fit closely enough the erroneous relation $r_A$ to $r_B$ when starting from $C$. However, we argue that using $A, C$ can help the model better identify the meaning of the frame in $t$. To confirm the benefit of $A, C$ on the performance, we define a new model in all points identical to the one defined in § 3.1, except $A, C$ do not appear in the input: we obtain $f_b(w_i|s, t, B)$ and $f_e(w_i|s, t, B)$. This new model can be seen as a simple transfer of $r_B$ from $s$ to $t$, instead of the analogical transfer we perform with the main model. The performance of this new model on the test data used for the analogical models is reported in the last three columns of Table 1. We observe a very significant drop in performance, with close to 19% for positive instances, with most of this gap transferred to "SR not found".

**Using different frames for $s$ and $t$.** In this section we study the applicability of our method when the frames containing the SRs in sentences $s$ and $t$ are different, in contrast to § 4.2 were we constrained our approach to same semantic frames.

Our starting intuition is that, as our model relies on semantic relations, if the frames of $s, t$ are different but semantically related, we should maintain high analogy solving performance. More specifically, the semantically closer the frames are, the higher the performance we should obtain.

The relations between frames indicated in FN1.7 (Baker et al., 1998; Baker, 2017) do not cover many frames, with a relation density[3] of the order of magnitude of $10^{-5}$ for all relations, except for *inheritance* which is closer to $10^{-4}$. To mitigate this effect and make manipulation more concise, we group the relations by meaning (we specify the FN1.7 relation and its inverse when relevant) and create denser, undirected relations:

- **Inheritance**: *Inherits from / Is Inherited by*;
- **Subframe**: *Subframe of / Has Subframe(s)*;
- **Causal and Temporal (C&T)**: *Precedes / Is Preceded by*, *Is Inchoative of*, and *Is Causative of*;
- **Other**: *See also*, *Uses / Is Used By*, and *Perspective on / Is Perspectivized in*.

To compute how related two frames are, we compute the smallest number of steps to reach one from the other following the relation[4]. If no path exists between two frames, we use the value "*not related*". We use 100 pairs of frames such that the two frames are different and neither appear in the training nor development data, which may overlap with the ones of the test sets of previous experiments. For each frame pair, we consider only SRs that are labeled the same in the two frames and generate up to 100 (positive) instances, for a total of 9834 instances.

We compute the Spearman correlation between the distance and the model accuracy, by considering each possible distance value as a class. We consider the case where no path exists between two frames (the vast majority of cases) as a separate class. The results are reported in Table 4, and the detailed accuracy and number of samples for each distance value for each relation is reported in Appendix C.4. There is however several limitations in our test method: firstly, if we exclude "No path", the test data contains only 1 distance value for *C&T* and 2 for *Subframe*, and each of these distance values is represented by only 100 instances; secondly, for *C&T*, *Subframe*, and *Other*, the correlation is much less significant than for *Inheritance*; and thirdly, we suspect that the relations we gathered in *Other* are too miscellaneous and not related enough to obtain meaningful relations between frames. All these limitations lead us to draw no conclusion with regard to the correlation between the performance and the relatedness in terms of the *C&T*, *Subframe*, and *Other* relations.

Nonetheless, for *Inheritance*, analysis of the Spearman's $\rho$ coefficient and the accuracy for each distance value (see Appendix C.4) indicates that

[3] The *density* of a relation between frames is number of pairs of frames that are related divided by the total number of frame pairs.

[4] This corresponds exactly to the node distance in the undirected graph of each relation.

the performance of the model increases for more closely related frames in terms of inheritance. In particular, for frames that are closely related, the performance is almost the same as when the sentences that activate the same frame (71.22% for distance of 2 against 72.21% when the frame is the same, and 49.49% for unrelated frames). These results indicate a certain flexibility of our approach with regards to the frame instantiated in the source sentence, which can help mitigate the scarcity of some frames.

### 4.2 Frame-semantic role labeling (FSRL)

Our analogical model can be used to propose frame annotations of unseen sentences. We assume a state of the art predicate identification and frame annotation method has been applied on the target sentence $t$ we want to annotate, providing us with the frame and the frame predicate. Relying on these first annotations, we create analogical equations to predict each of the remaining SRs. The difference with our general analogy solving setting in § 3.2 is that $A, C$ are the frame predicates of $s, t$. The source sentence $s$ is taken from our case base, *i.e.*, our training set. Our approach focuses on labeling SRs one by one, independently from each other. Repeating this operation for each SR of $F$ and solving the equation with our model allows us to get predictions for each SR of the frame in $t$. To demonstrate the feasibility of this approach, we apply our method on the test set of FN1.7, using source sentences from the corresponding training set. Note that our approach could be extended by using, the prediction of each SRs to improve and cross-check the predictions on the other SRs, as discussed in § 5.

**Model variants.** When implementing the approach, a key concern is the selection of the source sentence. We use two settings: (1) we use potentially different sources for each SR, or (2) we use the same source sentence for all the SR of the frame. As mentioned above, the basic use case for our approach is to label SRs one by one and independently. Setting (1) corresponds to this approach, where we are able to use the most fitting source for each SR. In this case our model achieves excellent results, outperforming the best state of the art model from (Lin et al., 2021) by a little under 4% under the best conditions (see Table 2). With setting (2) we want to see what happens when we present only one sentence to the system and get all the semantic roles out of this sentence using analogical transfer, with the advantage of reducing the number of sources to retrieve from the case base. Comparing the performance in (1) and (2) offers us bounds on the model performance with regards to the number of sentences used to label the SRs of a frame. Note that settings (1) and (2) do not cover all the SRs in the test set, as detailed further below. This is taken into account in the way we compute the F1, see Appendix D.2.

We determine additional bounds for the performance of the model by using several source selection algorithms. All our source selection algorithms are applied *a posteriori*, as we need to know the performance associated with each possible sources (a bit under a million analogies in total). To obtain the upper and lower limits of the performance of the model, we select the *best* and *worst* possible source in each setting. For the *best* source, we take any sentence that allows to successfully predict $e_t$ setting (1), while we take the sentence with the highest accuracy on the current frame in setting (2). We use a similar process for the *worst* source by obtaining the worst performance in each setting. To simulate the performance of the model in a realistic setting, we experiment with two *a priori* source selection algorithms: a naive *random* selection and a more involved selection based on *sentence similarity*. We approximate the *random* algorithm by averaging the accuracy over all possible sources for each SR. Our source selection based on *sentence similarity* is a proof of concept using the MiniLM[5] model (Wang et al., 2020) to obtain sentence embedding. We the apply the recommended *dot score*[6] to find the source most similar to the target among all possible sources. This sentence similarity model has two key limitations for our approach: it is not the state of the art model in term of semantic similarity, and it was not fine-tuned for its intended purpose of finding the most suited source. Its main purpose is to check the potential of such a source selection algorithm.

**FSRL performance.** The average performance over all the covered roles for the *best* and *random* selection is reported in Table 2, for settings (1)

[5] `all-MiniLM-L6-v2`, provided in the Sentence Transformers library (`https://www.sbert.net/docs/pretrained_models.html`) and recommended for its execution speed.

[6] For two embeddings $emb_1, emb_2$ and $\vartheta$ the angle between them, the dot score is the dot product $|emb_1||emb_2|cos(\vartheta)$. The more similar two embeddings are, the higher the dot score.

and (2). The *worst* and *sentence similarity* is reported in Appendix D.3. To compare the result of our approach and the state of the art we consider only the performance of the models when given the gold predicate and frame labels. For all models, an SR is correctly predicted if and only if the prediction is exactly the expected span. Our model has the potential to outperform the state of the art performance by at least 4%, if we manage to select the best possible source. Additionally, we expect our approach to be less sensible to failures in earlier stages of the frame semantic labeling process, as the performance of our approach does not degrade much when presented with mismatched but strongly related frames.

As mentioned above, the *best* performance obtained with our analogical approach is higher than the state of the art from (Lin et al., 2021) by close to 4% in setting (1). However, in setting (2) the F1 is 2.5% lower than the state of the art, despite the very close accuracy between (1) and (2) when only covered analogies are considered, as reported Appendix D.3. From these two results, the number of instances that are not covered by using a single source is a major limitation, while the performance on covered sentences does not degrade by a large margin. For instance, with setting (1) 95.25% of all SRs are covered[7], and 93.13% of frames have all their SRs covered. With setting (2) however, 84.78% of all SRs and 91.27% of frames are covered. The proof of concept *sentence similarity* selection achieves around 50% accuracy and does not bring significant improvements on randomly selecting a source. However, when studying the distribution of the highest sentence similarity for each source-target pair, only a small portion of target sentences have a very similar source available (1.34% of SRs above 0.7, see appendix Fig. 2). If we limit the study to these highly similar sources, we reach 70.45% of accuracy, only 2% under the best model from (Lin et al., 2021). We also found a significant correlation between the sentence similarity and the FSRL accuracy, with a Pearson-$r$ coefficient of $0.8997$ ($p$-value of $6.73^{-5}$) with slices of 0.1 on the similarity (see also appendix Fig. 3). The gap between the *best* and *worst* performance further highlights the importance of a sound source selection process.

| Source for the SRs | Different (1) | Same (2) |
|---|---|---|
| Random source | 47.32% | 45.18% |
| Best source | **76.08%** | 69.72% |
| Best from Lin et al. (2021) | 72.22% | |

Table 2: FSRL F1 (given the gold frame and predicate) on the FN1.7 test dataset for our approach when the *same* and *different* sources are used for a frame, as well as the best model from Lin et al. (2021).

**Performance for core and non-core SRs.** Our analogical training set (directly built from the FN1.7 ontology) covers only core SRs of a subset of all frames (22.38% of the frames of the test set). In Appendix D.5, we report the performance and the number of SRs depending on whether the corresponding frame is in the analogical training data and whether it is a core SRs. We notice a significant difference in performance between core and non-core SRs which can be expected as only core SRs were seen in training, however we suspect that part of this difference is due to the more subtle semantic link between non-core SRs and the predicate. We also notice that performance does not differ significantly between frames seen during training and unseen ones, highlighting the ability of our approach to generalize to unseen frames.

## 5 Discussion and perspectives

We propose an analogy dataset based on FN1.7 that complements the traditional datasets of analogies between words on factual and lexical semantics. Indeed, to study the SR of a group of words we need to consider contextual information, which is not that important a concern for factual or lexical semantics. By providing a clear definition of the underlying relation manipulated in the analogies, we also provide new insights on the study of semantic analogies between and within sentences. Building upon this dataset and using a very simple methodology, we propose an analogy solving approach that achieves high performance and is able to identify many unsolvable analogies. This approach also displays what can be seen as a tolerance with regards to mistakes in the analogical equation. Firstly, while ideally the sentences activate the same frame, the model maintains high performance for related but distinct frames. Secondly, the formulation of our model requires the first and third SR to be identical, but we show close performance when this rule is not respected. In

[7] An SR of the test set is not covered none of the sentences in the training set activate the corresponding frame or if the SR is never instantiated for this frame in the training sentences.

further work, we will explore how this tolerance improves performance with regards to mistakes in predicate and frame identification.

**Sentence selection.** Our experiments on FSRL provide us with bounds on the performance of our model with regards to the choice of the source sentence. Considering that the upper bound of our approach is, to the best of our knowledge, significantly higher than the state of the art, there is substantial potential in our analogy-based approach. This is especially true given that our approach is derived from a simple Ex-QA model, that we will improve in further work, for example by using a more involved architecture based on the principles of analogy. However, our experiments also show the importance of the selection of the source to achieve the best model performance, which will be the focus of further research.

In particular, with our experiments with enforcing a single source sentence for the SR of a frame, we identify that the main challenge in using the same source for all SR. Consequently, in further work we will explore a compromise approach, using a few sources prototypical sentences of each frame in order to cover as many SR of the frame as possible. This would allow to use a few carefully annotated sentences for each frame and significantly reduce the number of possible sources. Additionally, it is likely that using an ensemble of sources for each prediction would improve our model performance, but this involves significant exploration on the selection of sources and the aggregation of the predictions, so we prefer to dedicate future work to this specific extension.

Our proof of concept model for sentence selection does not achieve significantly better performance than randomly selecting the source. However, we observed that using highly similar sentences to use as sources was promising, but the number of targets that have similar sources is much to low to achieve the upper bound of the performance we can obtain. Consequently, a short term direction to improve the performance of our approach is to improve on the sentence similarity approach. Indeed, the sentence embedding model used is not the state of the art in term of semantic similarity, so selecting a different model might improve the results. Additionally, we will explore the possibility of fine-tuning the model for finding the most suited source or using metric learning to learn a sentence similarity model dedicated for our task.

**Labeling the SRs independently.** Our current FSRL approach labels each SR independently. We envision an extension of our framework in that regard, to match the state of the art FSRL systems (Lin et al., 2021; Zheng et al., 2022) and make full use of our analogy solving model. Indeed, we could, for each SR, use the prediction for the other SRs in addition to the frame predicate to create analogies, and use those new analogies to cross-check the predictions. In particular, we propose a two step procedure: *(i)* apply the current version of the approach to get a first prediction for each SR, and *(ii)* compute analogies between each pair of SRs to get an extra level of prediction for each SR taking into account the other SRs. The second step could be applied optionally to measure consistency and improve prediction quality.

**Generalizing the approach.** In this paper we focus exclusively on FN1.7, and we will extend the scope of application to FN1.5 if possible and to other similar dataset, such as PropBank (Pradhan et al., 2022). Also, as mentioned in §3, we intend to extend our approach to other languages. Indeed, in the past decade there has been a focus on providing labled resources for languages beyond English, with among other the French (Djemaa et al., 2016) and Swedish (Dannélls et al., 2021) FrameNets. However, this effort is for the most part limited to languages with many speakers, and frame annotation remains difficult and costly, limiting the amount and variety of annotation. To tackle this issue, further work will be done to offer a tool for FrameNet-style annotation in languages for which few or no labeled data is available, by leveraging analogical transfer and the multilingual embedding model mBert.

## 6 Conclusions

In this article, we provide a new task for studying analogical proportions at the frame semantic level, manipulating groups of words from different sentences. Our experiments show that a simple model based on semantic embeddings is enough to solve frame-semantic level analogies with a high performance. We also demonstrate the potential of our analogy solving approach to tackling FSLR. The main limitation of our FSLR approach is finding a fitting source sentence to perform analogical transfer, which will be the focus of future work as discussed below.

## Ethics statement

To the best of our knowledge, the dataset that we use (FN1.7) does not contain any sensitive information. Refer to (Baker et al., 1998; Baker, 2017) for further information.

In its current state, our approach does not offer significant improvements regarding fairness concerns, mostly due to the nature of the data manipulated and the annotation performed. However, applying our approach on other types of semantic annotations carries the potential of predictions explainable using the source used.

To the best of our knowledge, our use of mBert and MiniLM match their intended use (see the model cards linked in Appendix B).

## Limitations

Our approach suffers from the same limitation as other approaches using Bert models. For instance, it is difficult to analyze the involvement of each element of the input text in the result, and it is difficult to know what information is contained in each element of the output. If possible, we will explore in future work methods to better separate the input, effectively reducing this issue. Additionally, large pre-trained transformer models are known to be sensitive to small details in the formatting of the input and output, Additionally, the quality of semantic information in large pre-trained transformer models is known to depend on which transformer layers are used (van Aken et al., 2019), and this kind of model is very sensitive to input encoding (Zervakis et al., 2022). We do not perform any ablation study in that regard.

While we limit ourselves to the best quality annotations from FN1.7, the SR annotations are produced by human annotators which, by definition, may make errors. Any faulty or missing annotation may negatively influence our model. For instance, is an SR is present but not annotated, the model will learn to ignore the SR. Also, we limit ourselves to the core FEs and the frame predicates for training the model, and using peripheral and extra thematic FEs may improve performance.

We only perform a single run of our model, so all results should be confirmed by additional trials.

Our approach has been only tested on a single dataset written in the English language. As we use a multilingual pre-trained embedding model, our approach should work on all languages covered by the embedding model. However, the performance of our approach is expected to scale with the performance of the pre-trained models on each language, meaning a lower performance on less represented languages. In the future we plan to expand our approach to more datasets and languages.

Our approach does not take into account split predicates (only takes the first part) as opposed to the most recent approaches to FSRL. However, this can be achieved by concatenating the bits of predicates when specifying $A, B, C$ in the model input, eventual adding a special ellipsis token to mark the concatenation.

## References

Stergos D. Afantenos, Tarek Kunze, Suryani Lim, Henri Prade, and Gilles Richard. 2021. Analogies between sentences: Theoretical aspects - preliminary experiments. In *European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, volume 12897 of *Lecture Notes in Computer Science*, pages 3–18. Springer.

Stergos D. Afantenos, Suryani Lim, Henri Prade, and Gilles Richard. 2022. Theoretical study and empirical investigation of sentence analogies. In *International Joint Conference on Artificial Intelligence - European Conference on Artificial Intelligence, Interactions between Analogical Reasoning and Machine Learning Workshop*, volume 3174 of *CEUR Workshop Proceedings*, pages 15–28. CEUR-WS.org.

Collin F. Baker. 2017. Framenet: Frame semantic annotation in practice. *Handbook of Linguistic Annotation*, pages 771–811.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *International Conference on Computational Linguistics*, volume 1, pages 86–90, Montreal, Quebec, Canada. Association for Computational Linguistics.

Dana Dannélls, Lars Borin, and Karin Friberg Heppin. 2021. *The Swedish FrameNet++ Harmonization, integration, method development and practical language technology applications*. John Benjamins Publishing Company, Amsterdam, Philadelphia.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pages 4171–4186. Association for Computational Linguistics.

Marianne Djemaa, Marie Candito, Philippe Muller, and Laure Vieu. 2016. Corpus annotation within the French FrameNet: a domain-by-domain methodology. In *International Conference on Language Resources and Evaluation*, pages 3794–3801, Portorož, Slovenia. European Language Resources Association.

Aleksandr Drozd, Anna Gladkova, and Satoshi Matsuoka. 2016. Word embeddings, analogies, and machine learning: Beyond king - man + woman = queen. In *International Conference on Computational Linguistics, Technical Papers*, pages 3519–3530, Osaka, Japan. The COLING 2016 Organizing Committee.

Dedre Gentner. 1983. Structure Mapping: A Theoretical Framework for Analogy. *Cognitive Science*, 7:155–170.

Anna Gladkova, Aleksandr Drozd, and Satoshi Matsuoka. 2016. Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn't. In *North American Chapter of the Association for Computational Linguistics, Student Research Workshop*, pages 8–15, San Diego, California. Association for Computational Linguistics.

Douglas Hofstadter and Emmanuel Sander. 2013. *Surfaces and Essences: Analogy as the Fuel and Fire of Thinking*. Basic Books.

Douglas R. Hofstadter. 2001. Analogy as the Core of Cognition. In *The Analogical Mind: Perspectives from Cognitive Science*, chapter 15, pages 499–538. The MIT Press, Cambridge, Massachusetts.

Yves Lepage. 2003. *De l'analogie rendant compte de la commutation en linguistique*. Habilitation à diriger des recherches, Université Joseph-Fourier - Grenoble I.

Omer Levy and Yoav Goldberg. 2014. Linguistic regularities in sparse and explicit word representations. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 171–180, Ann Arbor, Michigan. Association for Computational Linguistics.

ZhiChao Lin, Yueheng Sun, and Meishan Zhang. 2021. A graph-based neural model for end-to-end frame semantic parsing. In *Conference on Conference on Empirical Methods in Natural Language Processing*, pages 3864–3874, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Esteban Marquer, Safa Alsaidi, Amandine Decker, Pierre-Alexandre Murena, and Miguel Couceiro. 2022. A Deep Learning Approach to Solving Morphological Analogies. In *International Conference on Case-Based Reasoning Research and Development*, volume 13405 of *Lecture Notes in Computer Science*, pages 159–174. Springer.

Esteban Marquer and Miguel Couceiro. 2023. Solving morphological analogies: from retrieval to generation. *CoRR*.

Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In *International Conference on Learning Representations, Workshop*.

Tomás Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, volume 26, pages 3111–3119. Curran Associates Inc.

Henri Prade and Gilles Richard. 2021. Analogical proportions: Why they are useful in ai. In *International Joint Conference on Artificial Intelligence*, pages 4568–4576. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Sameer Pradhan, Julia Bonn, Skatje Myers, Kathryn Conger, Tim O'gorman, James Gung, Kristin Wright-bettner, and Martha Palmer. 2022. PropBank comes of Age—Larger, smarter, and more diverse. In *Joint Conference on Lexical and Computational Semantics*, pages 278–288, Seattle, Washington. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100, 000+ questions for machine comprehension of text. In *Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392. Association for Computational Linguistics.

Anna Rogers, Aleksandr Drozd, and Bofang Li. 2017. The (too many) problems of analogical reasoning with word vectors. In *Joint Conference on Lexical and Computational Semantics*, pages 135–148, Vancouver, Canada. Association for Computational Linguistics.

Rupesh K. Srivastava, Klaus Greff, and Jürgen Schmidhuber. 2015. Training very deep networks. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Oren Sultan and Dafna Shahaf. 2022. Life is a circus and we are the clowns: Automatically finding analogies between situations and processes. In *Conference on Conference on Empirical Methods in Natural Language Processing*, pages 3547–3562, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Swabha Swayamdipta, Sam Thomson, Chris Dyer, and Noah A. Smith. 2017. Frame-semantic parsing with softmax-margin segmental rnns and a syntactic scaffold.

Peter D. Turney. 2008. The Latent Relation Mapping Engine: Algorithm and Experiments. *Journal of Artificial Intelligence Research*, 33:615–655.

Betty van Aken, Benjamin Winter, Alexander Löser, and Felix A. Gers. 2019. How does bert answer questions? a layer-wise analysis of transformer representations. In *International Conference on Information and Knowledge Management*, page 1823–1832, New York, NY, USA. Association for Computing Machinery.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Annual Conference on Neural Information Processing Systems*.

Georgios Zervakis, Emmanuel Vincent, Miguel Couceiro, Marc Schoenauer, and Esteban Marquer. 2022. An analogy based approach for solving target sense verification. In *International Conference on Natural Language Processing and Information Retrieval*, Bangkok, Thailand.

Ce Zheng, Xudong Chen, Runxin Xu, and Baobao Chang. 2022. A double-graph based framework for frame semantic parsing. In *North American Chapter of the Association for Computational Linguistics*, pages 4998–5011, Seattle, United States. Association for Computational Linguistics.

Xunjie Zhu and Gerard de Melo. 2020. Sentence analogies: Linguistic regularities in sentence embeddings. In *International Conference on Computational Linguistics*, pages 3389–3400, Barcelona, Spain (Online). Association for Computational Linguistics.

# A Training setup

The models were trained on Intel® Xeon(R) W-11955M CPU @ 2.60GHz × 16, 32 GiB RAM, NVIDIA RTX A5000 Mobile.

Training took around 1h30 on those setting, without using more than 4GiB RAM.

# B Code, data, and models

Once the anonymity period is over, the implementation of the model and the data preparation process will be made available on GitHub and, as much as possible, the trained model and evaluation results will be made available through an open data repository.

Our code relies on Python 3.9, as well as the following libraries: PyTorch, Huggingface (Transformers and Datasets), Sentence Transformers (also called SBert), Scipy, and Pandas.

The pretrained models we use are the mBert model (Devlin et al., 2019) called `bert-base-multilingual-cased` (`https://huggingface.co/bert-base-multilingual-cased`) and the MiniLM (Wang et al., 2020) model `all-MiniLM-L6-v2` (`https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2`).

The whole analogy solving model contains 177 269 762 parameters (177 268 224 for mBert model itself, 1 538 for the single layer neural networks used to predict the start and end token of the span).

# C Details on the analogies between semantic roles

## C.1 Dataset of analogies between semantic roles

To train our model and explore its analogy solving performance, we extract sentence examples from the FN1.7 ontology to build analogies $A : B :: C : D$, where $A, B, C, D$ are either instances of core SRs or the frame predicate. For each frame, we gather up to 1000 sentences with the annotation status of either FN1_Sent, Finished_Initial, or Finished_Checked (the 3 annotation status of the highest quality according to the documentation of FN1.7).

**Data augmentation.** To integrate analogical knowledge in our model, we use a data augmentation process based on the *symmetry* and *central permutation* axioms of analogical proportions (Lepage, 2003), following previous work on data augmentation for analogy solving (Marquer et al., 2022; Marquer and Couceiro, 2023). For each pair of semantic role and each pair of sentences of a frame, we generate 8 analogies: $A : B :: C : D$, $A : C :: B : D$, $D : B :: C : A$, $C : A :: D : B$, $C : D :: A : B$, $B : A :: D : C$, $D : C :: B : A$, $B : D :: A : C$. We exclude from our study analogies where $A = B$ and $C = D$. Indeed, the corresponding analogical equation would become $A : A :: C : x$ and the solution $x = C$ can be found without needing to explore the semantic relations between the elements of the analogical equation, which could degrade the quality of the training of the model.

**Different kinds of instance.** As mentioned in §3.2, it is possible that some SRs of a given frame are not instantiated in a given sentence. To account for this, we consider **positive instances** with analogies that can be solved because $r_D$ is instantiated in $t$ as $D$, and **negative instances** where the $r_D$ is not instantiated in $t$ and the analogy cannot be solved.

We create analogies such that:

- **for positive instances:** we create one analogy for a pair of distinct SRs such that both SRs are instantiated in both sentences;
- **for negative instances:** we create one analogy for a pair of distinct SRs such that one SR is instantiated in both sentences ($r_A, r_C$) and the second is instantiated in only one of them ($r_C$);
- **for instances with $r_A \neq r_C$:** we create one analogy for a triplet of distinct SRs such that one SR is instantiated in both sentences ($r_B, r_D$), the second is instantiated at least in the first sentence ($r_A$), and the third is instantiated at least in the second sentence ($r_C$).

**Balancing the SRs.** To maintain a good balance in the SRs presented, we randomly select analogies using the following process:

1. for $i$ SR tuple (pairs for positive and negative instances, triplets for $r_A \neq r_C$)
2. for $j_1, \ldots, j_i$ sentence pairs per SR tuple;
3. you randomly take (without putting back) one pair of sentences from each SR tuple and generate the corresponding analogy;

4. once a SR tuple does not have any remaining sentence pair, we exclude it;

5. we repeat steps 3-4 until we reach the number of analogies we want or no sentence pair remains.

An amusing analogy to this process is to consider the histogram of the number of pairs of sentences for each SR tuple. Now, consider that the histogram is an empty water tank, that we want to fill with the a volume of water corresponding to the number of analogies we want. If we pour the volume of water in the tank, it will fill the tank in a balanced manner. If an area of the tank is not high enough to accommodate as much water as the other areas, it will be filled to the brim and the remaining water will spread in the other parts of the tank.

**Amount of data in each part of the dataset.** To make our training and development set, we select randomly 250 frames from the leaf frames (*i.e.*, not having any frame inheriting from them in the FN1.7 ontology). Similarly, we select 100 leaf frames used in both the analogical test set and the $r_A \neq r_C$ set. For our training and development set, we first take up to 1000 positive and 1000 negative instances per frame. Then, to make the development set, we randomly take out 1000 positive and 1000 negative instances without considering which frame they are from. For the test set and the set with $r_A \neq r_C$, we take 100 instances of each class for each frame.

In total, the training set contains 249000 positive and 199816 negative instances, and the development set contains 1000 positive and 1000 negative instances. The test set contains 10000 positive and 8030 negative instances. Finally, the $r_A \neq r_C$ set contains 7760 instances.

### C.2 Token Position Error Rate (TPER)

TPER is an approximation of the Word Error Rate (WER) using the token positions. Let $b_x, e_x$ be the expected start and end token position of $x$ in the answer, $\hat{b}_x, \hat{e}_x$ the model predictions. The TPER is as follows, where $X \Delta Y$ is the symmetric difference of the sets $X$ and $Y$, and $[i, j]$ is the set of all integer values from $i$ to $j$, both included:

$$TPER(b_x, e_x, \hat{b}_x, \hat{e}_x) = \frac{|[\hat{b}_x, \hat{e}_x] \Delta [b_x, e_x]|}{|[b_x, e_x]|}.$$

### C.3 Safe Word Error Rate (SafeWER)

Word Error Rate (WER) is a measure of the average number of prediction errors in text, normalized by the number of expected words. To handle the empty targets we have for our negative instances, we add +1 to the denominator and obtain what we coin SafeWER. It corresponds to the average number of words to modify (replace, add, or remove) to obtain the *gold SR*. The formula of SafeWER is summarized below, with $Add.$, $Supr.$, and $Repl.$ the number of additions, suppression, and replacement respectively. Thus, $Add. + Supr. + Repl.$ is the number of modifications of the words of the prediction to obtain the *gold SR*, and $Target$ is the number of words in the *gold SR*.

$$SafeWER = \frac{Add. + Supr. + Repl.}{Target + 1}. \quad (2)$$

For an example of the behavior of the SafeWER, in the sentence "Your photographs have been substituted by our experts." with $D$ ="Your photographs", the prediction "photographs have been" would give a SafeWER of 1 and a WER of 1.5. However, with the target $D'$ not instantiated, "photographs have been" would give a SafeWER of 3 and "Your photographs" a SafeWER of 2 while the WER is undefined.

In Table 3, we report the SafeWER for our analogy solving model and the non-analogical model.

| | Failed instances | Overall |
|---|---|---|
| *Analogical model (using A,B,C)* | | |
| Positive | 0.96 | 0.27 |
| Negative | 2.53 | 0.71 |
| All | 1.66 | 0.46 |
| $r_A \neq r_C$ | 0.86 | 0.25 |
| *Non-analogical model (using only B)* | | |
| Positive | 0.85 | 0.40 |
| Negative | 2.75 | 0.68 |
| All | 1.42 | 0.52 |

Table 3: Analogy solving SafeWER for the analogical and non-analogical models. Instances where $r_A \neq r_C$ are considered only for the analogical model and are not counted in the overall performance (*All*).

### C.4 Detailed accuracy and number of instances using different frames in the analogy

Here we detail the performance of the model when using $s, t$ activating different frames. For each value of the *shortest path length (SPL)* between the two involved frames, we specify the accuracy (in %) and the number of instances, ordered by increasing distance. “*No path*” is set as the farthest distance possible, as very loosely related frames (high distance value) are more related than unrelated frames. Table 4 summarizes the correlation of the the relatedness and the performance for all 4 of our relation groups.

|  | $p$-value | $\rho$ | Unique values |
|---|---|---|---|
| Inheritance | **5.18e-68** | **-0.1744** | 10 |
| C&T | 6.97e-03 | +0.0272 | 2 |
| Subframe | 2.36e-03 | -0.0307 | 3 |
| Other | 3.69e-03 | +0.0293 | 10 |

Table 4: Correlations (Spearman correlation coefficient $\rho$ and corresponding $p$-values) between the success rates and the distance between frames frames for each relation in the dataset, in the setting where the two sentences do not activate. For reference, we also report the number of unique distance values that appear in the test data for each relation, including “*not related*” as a separate value. Boldface indicates the most significant result.

- Inheritance (*Inherits from* / *Is Inherited by*):
  - 2: 71.22% (900 instances),
  - 3: 68.88% (1128 instances),
  - 4: 66.21% (1400 instances),
  - 5: 63.12% (800 instances),
  - 6: 58.00% (200 instances),
  - 7: 56.00% (300 instances),
  - 9: 66.00% (200 instances),
  - 10: 54.50% (200 instances),
  - 12: 53.00% (200 instances),
  - No path: 49.49% (4506 instances).
- Subframe of (*Subframe of* / *Has Subframe(s)*):
  - 2: 73.00% (100 instances),
  - No path: 57.92% (9734 instances).
- Causal and temporal (*Precedes* / *Is Preceded by*, *Is Inchoative of*, and *Is Causative of*):
  - 1: 25.00% (100 instances),
  - 2: 73.00% (100 instances),
  - No path: 58.26% (9634 instances).
- Other (*See also*, *Uses* / *Is Used By*, and *Perspective on* / *Is Perspectivized in*):
  - 1: 63.67% (300 instances),
  - 2: 72.33% (300 instances),
  - 3: 26.42% (106 instances),
  - 4: 51.50% (200 instances),
  - 5: 42.50% (200 instances),
  - 7: 79.00% (100 instances),
  - 8: 70.50% (200 instances),
  - 10: 51.00% (100 instances),
  - 14: 12.50% (200 instances),
  - No path: 58.94% (8128 instances).

# D Details on the frame-semantic role labeling

## D.1 Encoding for the embedding model

The contextual embedding of each token from $t$ is computed by the mBert model accounting for $s, t, A, B, C$ by using the format shown in Fig. 1. It extends on the Ex-QA input format implemented in the HuggingFace library: “`[CLS] question [SEP] context [SEP]`”, where `[CLS]` and `[SEP]` are special tokens defined by mBert. To process multiple inputs of different length at the same time, a `[PAD]` padding token is added at the end of each input so that all the inputs have the same length. We add our own special tokens to indicate the boundaries of each element of our formulation to the transformer model: `[s]`, `[t]`, `[A]`, `[B]`, and `[C]`.

The *context*, which specifies where the answer should be found, corresponds to $t$ in our task. The *question* conditions the (semantic) content of the answer, and corresponds to $s, t, A : B :: C : x$ in our case. However, it is not necessary to provide $t$ in both the context and the question, so we limit the question to $s, A : B :: C : x$, resulting in what is displayed in Fig. 1.

$$\texttt{[CLS] [s] } \mathtt{w}_1^s \ldots \mathtt{w}_n^s \texttt{ [A] } \mathtt{w}_i^s \ldots \mathtt{w}_j^s \texttt{ [B] } \mathtt{w}_k^s \ldots \mathtt{w}_l^s$$
$$\texttt{[C] } \mathtt{w}_{i'}^t \ldots \mathtt{w}_{j'}^t \texttt{ [SEP] [t] } \mathtt{w}_1^t \ldots \mathtt{w}_m^t \texttt{ [SEP]}$$

Figure 1: Input format of the embedding model, where `[CLS]` and `[SEP]` are special tokens defined by mBert, to which we add `[s]`, `[t]`, `[A]`, `[B]`, and `[C]` to indicate the boundaries of each element of our formulation to the transformer model. On the first line we put the name of the object, for ease of teading, and on the second line we list the tokens to give a better idea of what the data looks like.

## D.2 Formula of F1 based on coverage

When considering the non-covered SRs as not predicted by our model, the formula for precision and recall become:

$$precision = \frac{\#\text{successfully predicted SRs}}{\#\text{covered SRs}}$$

$$recall = \frac{\#\text{successfully predicted SRs}}{\#\text{covered SRs} + \#\text{not covered SRs}}$$

### D.3 Extended results for source selection

We report in Table 5 the performance of all our sentence selection algorithm, as well as the best model from Lin et al. (2021).

| Score | Accuracy (covered) | | F1 | |
|---|---|---|---|---|
| Source | Same | Different | Same | Different |
| Best | 75.98% | 77.98% | 69.72% | **76.08%** |
| Sentence similarity | 53.08% | 52.32% | 48.71% | 51.05% |
| Random | 49.23% | 48.50% | 45.18% | 47.32% |
| Worst | 16.25% | 9.83% | 14.91% | 9.59% |

Table 5: FSRL performance (given the gold frame and predicate) when we consider potentially *different* sources for each SR of a frame and when we consider the *same* source for all SRs of a frame. We report F1 on the full dataset, while accuracy considers only SRs covered by each setting for better comparability.

### D.4 Source sentence similarity

In Fig. 2 we report the distribution of the highest sentence similarity for each source-target pair. We observe that 1.34% of SRs above 0.7. In Fig. 3 we report the performance of the model with regards to highest sentence similarity for each source-target pair.

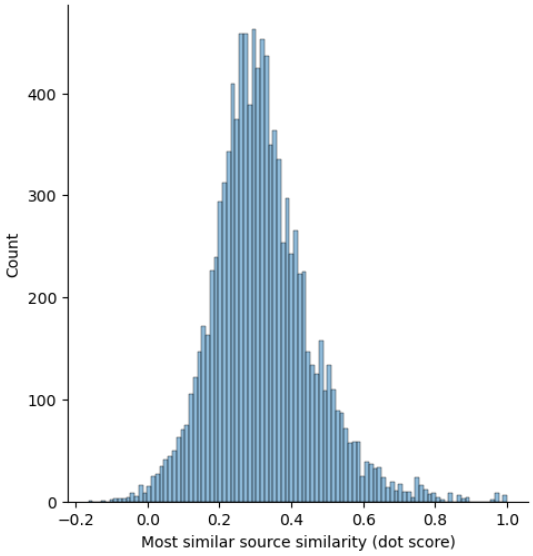

Figure 2: Source sentence similarity, in the setting where each SR can use different sources.

### D.5 Core SRs against non-core SRs

In Table 6, we report the performance and the number of SRs depending on whether the corresponding frame is in the analogical training data and whether it is a core SRs.

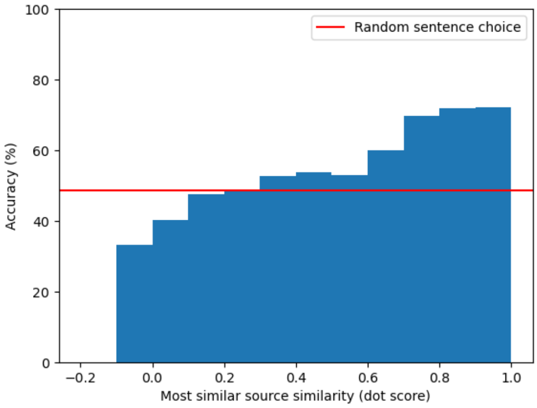

Figure 3: Source sentence similarity against FSRL accuracy, by bins of 0.1, in the setting where each SR can use different sources.

| | Frame seen in model training? | |
|---|---|---|
| | No | Yes |
| Non-core SR | 65.71% (1630) | 66.47% (170) |
| Core SR | 80.22% (6890) | 81.12% (2076) |

Table 6: FSRL accuracy on covered analogies (given the gold frame and predicate) in setting (1) using the best source, depending on whether the frame was seen in training and whether the SR is a core SR. The number of SRs in each category is reported in parenthesis.