FarFetched: An Entity-centric Approach for Reasoning on Textually Represented Environments

Anonymous ACL submission

Abstract

We address the problem of automatically acquiring knowledge from news articles and leverage it to estimate the veracity of a user's claim based on the supporting or refuting content within the accumulated evidence. We present *FarFetched*, an entity-centric approach for reasoning based on news, where latent connections between events, actions or statements are discovered via their identified entity mentions and are represented with the help of a knowledge graph. We propose a way of selecting specific subsets from the accumulated wealth of information based on the user hypothesis and construct relevant premises relying on the semantic similarity between them. We leverage textual entailment recognition to provide a measurable way for assessing whether the user claim is plausible based on the selected evidence. Our work is demonstrated on the less-resourced Greek language and supported by the training of state-of-the-art models for STS and NLI that are evaluated on benchmark datasets.

1 Introduction

In recent years, research in natural language understanding and textual inference has progressed significantly, leading in powerful models that can read, understand and reason about texts, reaching or even exceeding human performance. While these models are impressive as standalone achievements, they are either built following a closed-world assumption or require the supporting information to be provided along with the claim at hand in order to assert its validity. In the meantime, the global information explosion has led to an ever-expanding world of data that needs to be filtered, reviewed, analyzed, and processed for relevance and strategic significance. The challenges following the rapid increase in the amount of published information by news websites, RSS feeds, blogs and social media also affect commonsense reasoning tasks, as the arrival of new information may weaken or retract our initially supported inference, if taken into account.

The goals and contributions of this work[1] are: a) to formalize, develop and demonstrate a reasoning approach based on textual information from the continuous monitoring of news websites, where the user is able to input a claim in free text and assess its veracity in a measurable way and b) to train, evaluate and share[2] SotA models for the STS and NLI downstream tasks for the Greek language that support the core functionalities of our method.

2 Related Work

We are not aware of any work that attempts to perform the task of receiving an arbitrary user input as a hypothesis and assess its likelihood based on the accumulated knowledge from news articles (e.g. assess the likelihood of user input statements regarding a future event based on previous ones). Recent advances in the field of *event-centric* NLP introducing techniques for information extraction and event representation, machine comprehension and event prediction as well as knowledge acquisition at different levels of abstraction are briefly mentioned below.

Event representation methods usually leverage narrative event chains (Chambers and Jurafsky, 2008), knowledge graphs (Tang et al., 2019), question answer (QA) pairs (Michael et al., 2018) or event network embeddings (Zeng et al., 2021) to capture connections among events in a global context. These techniques are usually coupled with information extraction (IE) methods for joint entity, relation and event extraction (Lin et al., 2020) either on sentence-level (Kolluru et al., 2020) or on document level (Li et al., 2021), also covering cross-domain (Huang et al., 2018) and multi-

[1] FarFetched source code available here: [Link omitted for anonymity]

[2] Produced models available here: [Link omitted for anonymity.]

lingual (Papadopoulos et al., 2021) cases.

With regard to understanding relations between events and predicting future ones, recent trends include the temporal modelling of such problems with the help of narrative generation systems (Granroth-Wilding and Clark, 2016), attention-based prediction of event goals (Chen et al., 2020) and temporal knowledge graph embeddings (TKGE) (Zhu et al., 2021). Such approaches are usually demonstrated on close-domain problems, e.g. stock market prediction (Wu, 2020) or medical use cases (Deznabi et al., 2021).

The conversion of raw information derived from various sources into commonsense knowledge has also been established as a related line of work, with symbolic and neural approaches trying to resolve temporal and causal commonsense understanding of events (Hwang et al., 2021), or aiming to construct large-scale eventuality knowledge bases (Krzywicki et al., 2018) (Zhang et al., 2020).

Our method significantly differs from the aforementioned lines of work, as it relies on an *entity-centric* approach instead, where the identified entities are used as connectors between events, actions, facts, statements or opinions, thus revealing latent connections between the articles containing them. This is supported by semantic textual similarity and textual entailment recognition methods, ultimately aiming to decide whether a claim (hypothesis) provided by the user follows the existing evidence. A similar approach has been proposed for combining world knowledge with event extraction methods to represent coherent events, but relies on causal reasoning to generate plausible predictions of future events (Radinsky et al., 2012). A QA-based method for event forecasting is also relevant, but requires the accompanying news source to be provided along with the user's question (Jin et al., 2021). The latest advances regarding the technological concepts which comprise our methodology are provided below.

Entity linking (EL) is considered essential in many natural language understanding (NLU) systems, since it resolves the lexical ambiguity of entity mentions and determines their meanings in context. Typical EL approaches aim at identifying (named) entities in mention spans and linking them to entries of a KG (e.g. Wikidata, DBpedia) thus resolving their ambiguity. Recent methods combine the aforementioned tasks using local compatibility and topic similarity features (Delpeuch, 2019), pagerank-based wikification (Brank et al., 2017a) or neural end-to-end models that jointly detect and disambiguate mentions with the help of context-aware mention embeddings (Kolitsas et al., 2018).

The recent interest in sentence encoders for encoding diverse semantic sentence features into fixed-size vectors (Conneau et al., 2017) has resulted in SotA systems for Semantic Textual Similarity (STS) that are based on supervised cross-sentence attention (Raffel et al., 2020), Deep Averaging Networks (DAN) for sentence encoding (Cer et al., 2018) or siamese and triplet BERT-Networks (Reimers and Gurevych, 2019) to acquire meaningful sentence embeddings that can be compared using cosine-similarity.

Finally, the task of Natural Language Inference (NLI) -also known as Recognizing Textual Entailment (RTE)- can be used to investigate reasoning over long texts (a pair of premise and hypothesis phrases) into three classes: contradiction, entailment and neutral. The current state-of-the-art on this field relies on Transformer-based variants with global attention mechanisms (Beltagy et al., 2020), autoregressive language models for capturing long-term dependencies (Yang et al., 2019) and denoising autoencoders (Lewis et al., 2020).

3 Method

3.1 Problem Definition

Given a user input statement in free text (claim, hypothesis), we tackle the problem of deciding whether this statement is plausible based on the currently accumulated knowledge from news feeds. We also acknowledge the problem of constructing a relevant premise by analysing the wealth of information contained in hundreds of millions of articles that inevitably creates a poverty of attention and the need to devise an efficient way for extracting only contextually and semantically relevant text subsets to verify or refute the user's hypothesis. While our work does not primarily focus on better sentence embeddings and natural language inference techniques, we also target the lack of such models for the Greek language.

3.2 Our approach

FarFetched combines a series of offline (performed periodically) and online (performed upon user input) processes to crawl for news articles, annotate their context with named entities and derive a rel-

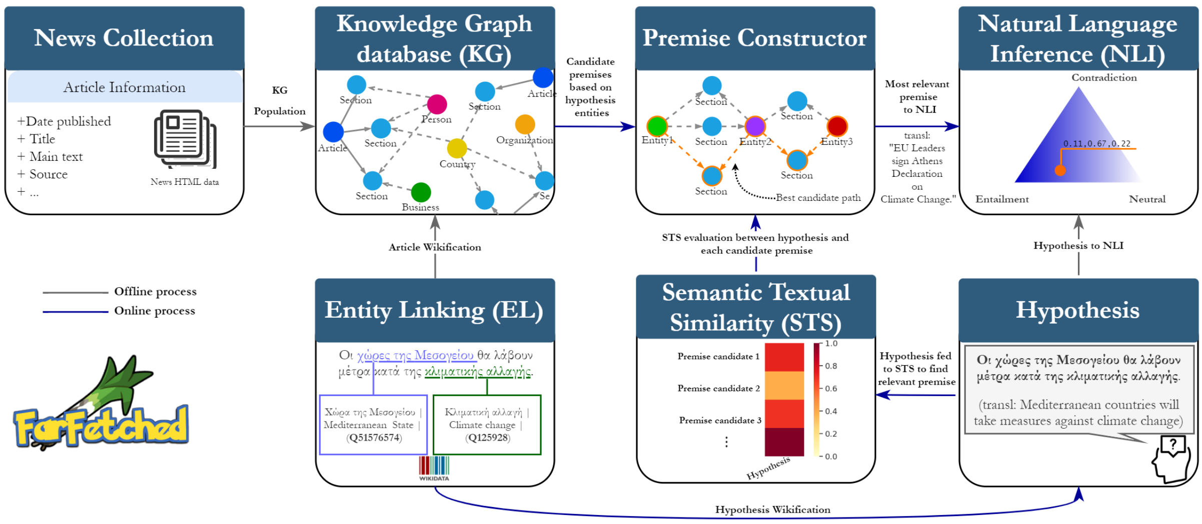

Figure 1: Overview of the FarFetched approach

evant subset of the stored content to reason about the validity of the user hypothesis in an NLI setting. These operations are visualised in Figure 1, can be summarized as follows and are described more thoroughly in the following subsections:

- **Offline processes:**

News Collection: A news crawler is deployed to accumulate information by extracting HTML content from news websites.

KG Database population: The crawled content (article title, text sections, publication date etc.) is processed and stored in a knowledge graph allowing a more structured representation.

Entity Linking (on articles): Wikification is applied to each article section to identify concepts and link events based on their disambiguated entity mentions.

- **Online processes:**

Entity Linking (on hypothesis): Upon user input, the Entity Linking process will annotate the hypothesis, aiming at finding entities that can be linked with those in the KG database.

Premise Constructor: The identified entities of the previous phase serve as the starting point for the construction of a contextually and semantically relevant premise. The constructor returns all the article sections that connect the identified entities using a shortest path approach.

Semantic Textual Similarity: This process aims at selecting the best premise by comparing the vector representations of the hypothesis with each candidate premise in terms of semantic similarity.

Natural Language Inference: The best candidate premise and the hypothesis are fed into an NLI model that determines whether the latter is entailed, contradicted or neutral to the former and outputs the probability scores for each case.

3.2.1 News Collection

News articles are collected via the *news-please* framework (Hamborg et al., 2017), a multi-language, open-source crawler and extractor for heterogeneous website structures. It is capable of extracting the major elements of news articles (i.e., title, lead paragraph, main content, publication date, author, etc.), featuring full website extraction and requiring only the root URL of a news website to crawl it completely. The framework relies on *scrapy* (Kouzis-Loukas, 2016) to download each article's HTML and supports 4 different modes of operation in order to find all articles published by a news outlet : i.RSS feed analysis, ii.recursive crawling by following internal links, iii.sitemap analysis for fetching articles from the whole website and iv.automatic mode, which prioritizes sitemap analysis and falls back to recursive crawling in case of error.

3.2.2 KG Database Population

The crawled news articles are used to populate a Knowledge Graph (KG) database. In order to store article-related information, the open-source version of the *Neo4j* graph database management system (Webber, 2012) was used, as it supports native graph storage and processing functionalities along with a convenient browser visualization tool. In its raw form, the KG database contains only two types of nodes with their respective properties:

Articles, which represent crawled news articles and Sections that represent the sentences of each article's main text (concatenated title and article body). Each Article node is linked to one or more Section nodes by the HAS_SECTION relationship. The KG is enriched with additional entities and relationships via the Entity Linking process.

3.2.3 Entity Linking

Given that our approach relies on largely unstructured textual documents that lack explicit semantic information, Entity Linking (EL) constitutes a central role in revealing latent connections between seemingly uncorrelated article sections. To this end, *FarFetched* employs a type of semantic enrichment and entity disambiguation technique known as wikification, which involves using Wikipedia concepts as a source of semantic annotation. We call the JSI Wikifier service (Brank et al., 2017b), a free Web API with multilingual support to annotate both the content of the crawled news articles (offline process) and the hypotheses received by the user (online process). The service applies pagerank-based wikification on input text to identify phrases that refer to entities of the target knowledge base (Wikipedia) and return their corresponding Wikipedia URL and WikiData entity ID, which is a number prefixed by a letter. The latter is used as a unique identifier for storing these entities as Entity nodes back to our KG database and linking them with the crawled article Section nodes, resulting to a more tightly connected graph, where each entity/concept is connected to multiple sections via the HAS_ENTITY relationship.

3.2.4 Premise Constructor

In a typical NLI setting, a premise represents our knowledge about an event and is used to infer whether a relevant hypothesis follows from it or not. In our case, however, multiple independent descriptions of the same or similar events (i.e. multiple news articles focusing on the same entities) might be available. It would therefore be beneficial to leverage the linked article sections of our KG by involving inference over longer premise texts and aggregation of information from multiple candidate premise sentences (Lai et al., 2017). Our approach is simple and comprises the following steps (online process):

1. The user inputs a free-text statement in Greek which serves as a hypothesis, e.g. transl: "*Mediterranean countries will take measures against climate change*" .

2. The hypothesis is passed through the Entity Linking phase (wikification) and one or more Wikidata concepts are identified as entity IDs (e.g. Q41, Q51576574).

3. The KG is queried for the aforementioned concepts (represented as Entity nodes) and tries to find all possible shortest paths between them. Given the implemented graph structure and a sequence of n alternating Entity-Section nodes where each Section node is connected to at least one Entity node, this translates to a minimum path length of $2(n-1)$.

4. For each sequence, the textual information contained in all Section nodes is concatenated to form a candidate premise. Their relevance with the hypothesis at hand is assessed during the Semantic Textual Similarity phase.

3.2.5 Semantic Textual Similarity

We apply a sentence embeddings method to extract and compare the vector representations of the user's hypothesis and each candidate premise, in order to select the most semantically relevant candidate for the final NLI phase. Despite the abundance of multilingual language models (e.g. m-BERT, XLM) that cover most common languages, a pretrained multilingual sentence embeddings model does not generally perform well in downstream tasks for less-resourced languages like Greek (Koutsikakis et al., 2020). Furthermore, given that the vector spaces between languages are not aligned, sentences with the same content in different languages could be mapped to different locations in the common vector space. To overcome this obstacle, we followed a multilingual knowledge distillation approach proposed by Reimers and Gurevych, 2020 to train a Greek sentence embedding model using parallel EN-EL (English-Greek) sentence pairs using the *sentence-transformers* library (Reimers and Gurevych, 2019). Our Greek student model (*XLM-roberta-base*) was trained using the parallel pairs to produce vectors for the EN-EL sentences that are close to the teacher's pretrained English model (*distilrobertabase-paraphrase-v2*) ones. Using the trained model, we are able to compare the produced vector representations between the hypothesis and each candidate premise in terms of Semantic Textual Similarity (STS) using the cosine similarity metric and forward the best candidate premise to the last phase (NLI) of the online process.

3.2.6 Natural Language Inference

The last step of our process relies on Natural Language Inference (NLI) to determine whether the hypothesis is true (entailment), false (contradiction), or undetermined (neutral), given the most relevant premise of the previous phase. To tackle the aforementioned multilinguality issues of pretrained language models on low-resource languages, we finetuned a Greek *sentence-transformers* Cross-Encoder (Reimers and Gurevych, 2019) model (*XLM-roberta-base*) for the NLI task. The model was trained on the Greek and English version of the combined SNLI (Bowman et al., 2015) and MultiNLI (Williams et al., 2018) corpora (AllNLI). To create the Greek version of AllNLI, the English-to-Greek machine translation model by Papadopoulos et al., 2021 was used[3]. The trained model takes the premise-hypothesis pair as input and predicts one of the following labels for each case: "contradiction": 0, "entailment": 1, "neutral": 2. The logits for each class are then converted to probabilities using the softmax function. These labels along with their probability scores can be used to assess whether the user statement is verified by the accumulated knowledge on our KG Graph database.

4 Experiments

4.1 Setup

The technical details for each building block of *FarFetched* are provided below:

News Collection and Storage: The python package of *news-please*[4] was used to create an initial corpus of news articles to support our experiments. We used the automatic mode on the root URLs of two popular Greek news sites in order to recursively crawl news from a diverse topic spectrum, spanning from 2018 until 2021. We collected 13,236 articles, containing 31,358 sections it total. The *Neo4j Community Edition v4.3*[5] was used to store the crawled articles and sections as nodes and create their in-between relationships.

Entity Linking: A Python script producing POST requests to the free web API of *JSI Wikifier*[6] was used to annotate the article sections and enrich the KG with Wikidata entities. A total of 2,516 Wikidata entities of different types (e.g. sovereign states, cities, humans, businesses, organizations, academic institutions etc.) were identified in the crawled articles. A pageRankSqThreshold of 0.80 was set for pruning the annotations on the basis of their pagerank score.

Premise Constructor: To create candidate premises, a parametrizable Cypher query executed via a Python script is used that takes the identified entities in the hypothesis as parameters and returns the concatenated article sections that link these entities together. For our experiments, the maximum number of relationships between the alternating Sections and Entities was set to $2(n-1)$ (shortest path), while the script returns at most 50 candidate premises in descending order based on path length. These parameters can be modified if longer premise candidate sets of sentences are required.

Semantic Similarity: The *sentence-transformers*[7] library was used to finetune a bilingual (Greek-English) *XLM-roberta-base* model (~270M parameters with 12-layers, 768-hidden-state, 3072 feed-forward hidden-state, 8-heads) using 340MB of parallel (EN-EL) sentences from various sources (e.g. OPUS, Wikimatrix, Tatoeba). The model was trained for 4 epochs with a batch size of 16 on a machine with a single NVIDIA GeForce RTX3080 (10GB of VRAM).

Natural Language Inference: Using the above hardware setting, a Cross-Encoder *XLM-roberta-base* of the same architecture was finetuned on the Greek-English AllNLI dataset with the *sentence-transformers* library. The model was trained for a single epoch, using a train batch size of 6 due to memory constraints.

4.2 Main results

In this section we perform a qualitative demonstration of *FarFetched*'s overall performance and also provide quantitative results for our two trained models with regard to common benchmark datasets.

4.2.1 Demonstration of the overall methodology

Given the particularity of the *FarFetched* approach and the specific nature of its goals, it is difficult to

[3] Greek AllNLI version available here: [Link omitted for anonymity]

[4] https://github.com/fhamborg/news-please

[5] https://neo4j.com/download-center

[6] http://wikifier.org/

[7] https://www.sbert.net

assess its performance quantitatively via a benchmark dataset. To this end, we provide a set of examples that aim at showcasing the capabilities of our system in deciding about the validity of a user's hypothesis based on the accumulated information. These scenarios include two parts each and are shown in Table 1. The original data (in Greek) are available in the Appendix[8].

In *Scenario 1*, two contradicting user hypotheses with the same entity mentions are provided by the user. Given that they refer to the same entities, the system fetches the same candidate premises pool for each case in order to evaluate their validity. The most relevant one (score in bold) is selected for the NLI phase, where the verdict is that the premise entails the first hypothesis (1a) and contradicts the second (1b).

In *Scenario 2*, we investigate the sensitivity of our approach in exploiting new information to evaluate a hypothesis. The hypothesis of 2a triggers the premise constructor which returns multiple candidate premises, the most relevant of them having a similarity score of 0.6665. During the NLI evaluation phase, the verdict is entailment, but with low confidence. In 2b the same hypothesis is evaluated, but with the addition of an artificial news section that is clearly more relevant to the claim at hand. This is successfully identified by *FarFetched*'s STS component which selects the new section as the best candidate, providing a more confident NLI decision. This shift in NLI verdict is visualized in Figure 2. Given that *FarFetched* can provide reasoning on the constantly updated news flow, monitoring such shifts could be useful for identifying trend changes, especially for cases that benefit from long-term planning (business, market, politics etc.)

Scenario 3 is similar to 2, as the same hypothesis is evaluated on the existing candidate premises pool (3a) and on an artificial section added in 3b. However, in this case the added information is an excerpt from a person's interview. While our approach correctly identifies the relevance of this section to the user hypothesis affecting the NLI decision, there is no way of knowing whether this claim is truthful or not. This is discussed in more detail in Section 5.

[8]omitted for camera-ready version according to guidelines. See supplementary material.

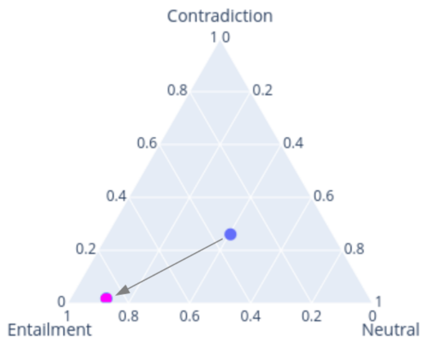

Figure 2: Shift in NLI verdict from Scenario 2a (blue) to 2b (pink) of Table 1.

4.2.2 STS performance

The performance of our sentence embeddings model was evaluated on the test subset of the Semantic Textual Similarity (STS) 2017 dataset (Cer et al., 2017). Given that the original STS2017 dataset does not provide sentence pairs in Greek, we manually created a cross-lingual version for the English-Greek pair with the help of a native speaker[9]. The performance is measured using Pearson and Spearman correlation between the predicted similarity score and the gold score. We also provide results in terms of translation matching accuracy by evaluating if the source and target language embeddings are close using cosine similarity. The results are shown in Table 2. We obtain a slightly better performance both in terms of STS and translation matching compared to the current state-of-the-art multilingual model by Reimers and Gurevych, 2019.

4.2.3 NLI performance

We benchmark our trained model on the Greek subset of the XNLI-test benchmark (Conneau et al., 2018) that contains 5,010 premise-hypothesis pairs. The results are shown in Table 3. Despite not having used the XNLI-train set for our training, we achieve a 1% gain over the multilingual XLM-R (Conneau et al., 2020) and are on par with the monolingual Greek-BERT by Koutsikakis et al., 2020. Given that our model was trained on a mixture of Greek and English sentence pairs, it is more suitable for corpora that also contain English terms (e.g. technology, science topics) without suffering from the under-representability of the Greek language occurring in multilingual models.

[9]EN-EL version of STS2017 dataset available here: [Link omitted for anonymity]

#	User Hypothesis	Fetched candidate premises (similarity)	NLI scores (c;e;n)
1a	Denmark and Austria believe that the European Union should increase aid to refugees.	Austria and Denmark also want to increase EU support for countries hosting refugees near crisis hotspots so that they do not travel to Europe. (**0.8505**)	0.014 ; **0.958** ; 0.028
1b	Denmark disagrees with Austria on the management of immigration issues in the European Union.	Checked by police at the Airport Police Departments ... the foreigners presented forged travel documents ... in order to leave the country for France, Germany, Italy, the Netherlands, Denmark, Spain and Norway. (0.2283)	**0.951** ; 0.002 ; 0.047
2a	The United States plans to impose sanctions on Iran.	Iran faces dilemma over whether to comply of Washington or will lead to collapse. The sanctions that came back in force today, will force the government of the Islamic Republic to accept the US claims regarding the Iranian nuclear program and Iranian activities in the Middle East East because, otherwise, the regime will be in danger to collapse, claimed Israel Kats, the Israeli minister responsible for Information Services. (**0.6665**) Why Greece was exempted from US sanctions on Iran. New US sanctions on oil exports from Iran have been in force since November 5. (0.6324) "We are always in favor of diplomacy and talks ... But the Conversations need honesty ... The US is pushing again sanctions on Iran and withdraw from the nuclear deal "(of 2015) and then they want to have conversations with us", Rohani said in a speech that was broadcast live on television. (0.5151) The condemnation of the banker Mehmet Atila is energizing the climate between the USA and Turkey. The already tense relations between Turkey and the USA are strengthened by the decision of the Manhattan federal court, which on Wednesday found Turkish banker Mehmet Hakan Attila guilty of participating in a conspiracy to offer help Iran to circumvent US financial sanctions. (0.4018)	0.220 ; **0.454** ; 0.326
2b		... + Following the collapse of the last talks between the US and Iran, the announcement of additional sanctions is expected in the coming days. (**0.7195**)	0.006 ; **0.952** ; 0.042
3a	Apple is trying to compete with Netflix in the production of television content.	Apple is expected to spend about $ 2 billion this year creating original content that it hopes will compete with Netflix, Hulu and Amazon, already established in the television audience. (**0.7107**)	0.004 ; **0.967** ; 0.029
3b		... + "We're not trying to compete with Netflix on TV," an Apple spokesman said in an interview. (**0.7134**)	**0.982** ; 0.008 ; 0.010

Table 1: Demonstration of FarFetched on 3 scenarios. All sentences are translated from Greek to English for better readability. The "+" sign denotes the addition of an artificial premise to the existing candidates for the same scenario to showcase the sensitivity of our approach in accumulating new information. Underlined hypothesis text indicates the entities annotated during the wikification process. The similarity scores of the candidate premises in bold signify the best candidate. Similarly, the NLI score in bold represents the probability of the predicted label (contradiction, entailment or neutrality respectively).

Model	STS2017		Translation Matching	
	r	**ρ**	**Acc. (eln2el)**	**Acc. (el2en)**
XLM-RoBERTa-base (Ours)	**83.30**	**84.32**	**98.05**	**97.80**
Paraphrase-multilingual-mpnet-base-v2 (UKP-TUDA)	82.71	82.70	97.50	97.35

Table 2: Comparison of our sentence-embeddings model in terms of Pearson (r) and Spearman (ρ) cosine similarity on the STS2017 set (EN-EL version) and in terms of translation matching accuracy.

Model	F1-score
XLM-RoBERTa-base ***(Ours)***	**78.3**
Greek-BERT (AUEB)	78.6 ± 0.62
XLM-RoBERTa-base (Facebook)	77.3 ± 0.41
M-BERT (Google AI Language)	73.5 ± 0.49

Table 3: Model comparison of our NLI model in terms of F1-score on the Greek subset of XNLI-test dataset.

5 Error Analysis

We acknowledge that *FarFetched* is possible to encounter errors in 3 main areas: entity linking, premise construction and entailment recognition (NLI). These are briefly addressed below.

Entity Linking: Highly ambiguous entities (e.g. "Washington" could refer to the US state or to "George Washington") and name variations (e.g. "European Union" and "EU") pose challenges to any entity linking method. Since we claim that our approach is entity-centric, a wrong annotation of the hypothesis' or article's entities will lead to irrelevant candidate premises and increase the probability of "neutral" NLI verdicts. Moreover, the tunable sensitivity of the JSI Wikifier implies a tradeoff between a precision-oriented and a recall-oriented strategy, the latter resulting in a richer KG, but also being prone to false-positive annotations.

Premise Construction: This initial version of our approach relies solely on the STS comparison between the premises that contain the same entities as the hypothesis, based on a shortest path approach discussed in Section 3.2.4. In cases where a larger number of entities are identified in the user hypothesis, finding the traversal path between the alternating Entity-Section nodes can be a time-consuming operation. Moreover, there is no guarantee that the shortest path is able to capture the optimal candidate premises; to this end an aggregation of the top n most relevant premises is considered as an alternative. Finally, there is currently neither a temporal evaluation of the candidate premises with regard to the hypothesis nor a distinction between opinions and facts; all candidates are treated as equal.

Natural Language Inference: Recognizing the entailment between a pair of sentences partially depends on the tense and aspect of the predications. Especially in our case, where we rely on information from news articles, tense plays an important role in determining the temporal location of the predication (i.e. in the past, present or future), while the aspectual auxiliaries signify an event's internal constituency (e.g. whether an action is completed or in progress). While the work of Kober et al., 2019 indicates that language models encode a substantial amount of morphosyntactic information regarding tense and aspect, they are unable to reason based only on these properties. To this end, user hypotheses with a high presence of such semantic properties should be avoided.

6 Conclusions

In this work, we presented a novel approach for reasoning based on the accumulated knowledge from the continuous ingestion and processing of news articles. *FarFetched* is able to evaluate the validity of any arbitrary human input (claim, hypothesis) in free text given the existing evidence, relying on the pillars of news crawling, knowledge graphs, entity linking, semantic textual similarity and natural language inference.

We showcased the effectiveness of our method in divese scenarios and acknowledged its weaknesses and limitations. As byproducts of our work, we trained and opensourced an NLI and a sentence embeddings model for the less-resourced Greek language, achieving state-of-the-art performance on the XNLI and STS2017 benchmarks respectively. While the implementation of our approach is focused on Greek, its modular architecture allows it to be repurposed for any language for which the corresponding models exist.

For future work, we intend to address some of the limitations of our method mentioned in Section 5, focusing primarily on an optimal setting for our entity linking component as well as on devising an improved strategy for constructing the candidate premises pool and evaluating their suitability.

References

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *CoRR*, abs/2004.05150.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Janez Brank, Gregor Leban, and Marko Grobelnik. 2017a. Annotating documents with relevant wikipedia concepts. *Proceedings of SiKDD*.

Janez Brank, Gregor Leban, and Marko Grobelnik. 2017b. Annotating documents with relevant wikipedia concepts. In *Proceedings of the Slovenian Conference on Data Mining and Data Warehouses (SiKDD 2017)*, pages 218–223.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo 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)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 169–174, Brussels, Belgium. Association for Computational Linguistics.

Nathanael Chambers and Dan Jurafsky. 2008. Unsupervised learning of narrative event chains. In *Proceedings of ACL-08: HLT*, pages 789–797, Columbus, Ohio. Association for Computational Linguistics.

Muhao Chen, Hongming Zhang, Haoyu Wang, and Dan Roth. 2020. What are you trying to do? semantic typing of event processes. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 531–542, Online. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark. Association for Computational Linguistics.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Antonin Delpeuch. 2019. Opentapioca: Lightweight entity linking for wikidata. *arXiv preprint arXiv:1904.09131*.

Iman Deznabi, Mohit Iyyer, and Madalina Fiterau. 2021. Predicting in-hospital mortality by combining clinical notes with time-series data. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4026–4031, Online. Association for Computational Linguistics.

Mark Granroth-Wilding and Stephen Clark. 2016. What happens next? event prediction using a compositional neural network model. *Proceedings of the AAAI Conference on Artificial Intelligence*, 30(1).

Felix Hamborg, Norman Meuschke, Corinna Breitinger, and Bela Gipp. 2017. news-please: A generic news crawler and extractor. In *Proceedings of the 15th International Symposium of Information Science*, pages 218–223.

Lifu Huang, Heng Ji, Kyunghyun Cho, Ido Dagan, Sebastian Riedel, and Clare Voss. 2018. Zero-shot transfer learning for event extraction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2160–2170, Melbourne, Australia. Association for Computational Linguistics.

Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2021. (comet-) atomic 2020: On symbolic and neural commonsense knowledge graphs. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(7):6384–6392.

Woojeong Jin, Rahul Khanna, Suji Kim, Dong-Ho Lee, Fred Morstatter, Aram Galstyan, and Xiang Ren. 2021. ForecastQA: A question answering challenge for event forecasting with temporal text data. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4636–4650, Online. Association for Computational Linguistics.

Thomas Kober, Sander Bijl de Vroe, and Mark Steedman. 2019. Temporal and aspectual entailment. In *Proceedings of the 13th International Conference on Computational Semantics - Long Papers*, pages 103–119, Gothenburg, Sweden. Association for Computational Linguistics.

Nikolaos Kolitsas, Octavian-Eugen Ganea, and Thomas Hofmann. 2018. End-to-end neural entity linking. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 519–529, Brussels, Belgium. Association for Computational Linguistics.

Keshav Kolluru, Samarth Aggarwal, Vipul Rathore, Mausam, and Soumen Chakrabarti. 2020. IMoJIE:

Iterative memory-based joint open information extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5871–5886, Online. Association for Computational Linguistics.

John Koutsikakis, Ilias Chalkidis, Prodromos Malakasiotis, and Ion Androutsopoulos. 2020. Greek-bert: The greeks visiting sesame street. In *11th Hellenic Conference on Artificial Intelligence*, SETN 2020, page 110–117, New York, NY, USA. Association for Computing Machinery.

Dimitrios Kouzis-Loukas. 2016. *Learning Scrapy*. Packt Publishing Ltd.

Alfred Krzywicki, Wayne Wobcke, Michael Bain, Susanne Schmeidl, and Bradford Heap. 2018. A knowledge acquisition method for event extraction and coding based on deep patterns. In *Knowledge Management and Acquisition for Intelligent Systems*, pages 16–31, Cham. Springer International Publishing.

Alice Lai, Yonatan Bisk, and Julia Hockenmaier. 2017. Natural language inference from multiple premises. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 100–109, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Sha Li, Heng Ji, and Jiawei Han. 2021. Document-level event argument extraction by conditional generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 894–908, Online. Association for Computational Linguistics.

Ying Lin, Heng Ji, Fei Huang, and Lingfei Wu. 2020. A joint neural model for information extraction with global features. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7999–8009, Online. Association for Computational Linguistics.

Julian Michael, Gabriel Stanovsky, Luheng He, Ido Dagan, and Luke Zettlemoyer. 2018. Crowdsourcing question-answer meaning representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 560–568, New Orleans, Louisiana. Association for Computational Linguistics.

Dimitris Papadopoulos, Nikolaos Papadakis, and Nikolaos Matsatsinis. 2021. PENELOPIE: Enabling open information extraction for the Greek language through machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 23–29, Online. Association for Computational Linguistics.

Kira Radinsky, Sagie Davidovich, and Shaul Markovitch. 2012. Learning to predict from textual data. *Journal of Artificial Intelligence Research*, 45:641–684.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

Jizhi Tang, Yansong Feng, and Dongyan Zhao. 2019. Learning to update knowledge graphs by reading news. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2632–2641, Hong Kong, China. Association for Computational Linguistics.

Jim Webber. 2012. A programmatic introduction to neo4j. In *Proceedings of the 3rd Annual Conference on Systems, Programming, and Applications: Software for Humanity*, SPLASH '12, page 217–218, New York, NY, USA. Association for Computing Machinery.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Xianchao Wu. 2020. Event-driven learning of systematic behaviours in stock markets. In *Findings* *of the Association for Computational Linguistics: EMNLP 2020*, pages 2434–2444, Online. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Qi Zeng, Manling Li, Tuan Lai, Heng Ji, Mohit Bansal, and Hanghang Tong. 2021. GENE: Global event network embedding. In *Proceedings of the Fifteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-15)*, pages 42–53, Mexico City, Mexico. Association for Computational Linguistics.

Hongming Zhang, Xin Liu, Haojie Pan, Yangqiu Song, and Cane Wing-Ki Leung. 2020. Aser: A large-scale eventuality knowledge graph. In *Proceedings of The Web Conference 2020*, WWW '20, page 201–211, New York, NY, USA. Association for Computing Machinery.

Cunchao Zhu, Muhao Chen, Changjun Fan, Guangquan Cheng, and Yan Zhang. 2021. Learning from history: Modeling temporal knowledge graphs with sequential copy-generation networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(5):4732–4740.