A Semantic Frame Graph for Information Extraction

Michał Gałusza
Faculty of Cybernetics, Military University of Technology, Warsaw, Poland

Keywords: Information Extraction, Relationship Detection, Natural Language Processing, Graphical Representation of Text.

Abstract: The following paper describes a graphical representation of a short text based on the semantic frames (the Semantic Frame Graph) using Semantic Role Labeling (SRL). It can be a foundation of alternative approach for open information extraction (OIE). The approach postprocesses the output of pretrained SRL classifier and it does not use complex rules, training sets nor significant corpus to decompose sentences. Proposed decomposition and representation reduces number of paths between entities dropping ones that are linguistically unmotivated, generates sequences of frames as paths which can be controlled using dialog coherence approach which further increases plausibility of semantic relationship between entities.

1 INTRODUCTION

Information Extraction (IE) is a process of converting unstructured information held in text into a structured, for example, graphical representation as in knowledge graphs (KG). The KGs represent information as typed relations between entities, which are fundamental for developing methods that have the potential for sophisticated reasoning (Ji, 2021).

Open Information Extraction (OIE) paradigm aims to extract all possible semantic relationships between entities available regardless of the genre of text (financial, medical or technical). This is achieved without dedicated, supervised training of anticipated relationships. OIE systems, however, are not fully unsupervised. They use a collection of patterns, seeding samples, predefined heuristics, scoring functions and distant supervision (Fader, 2011),(Banko, 2007),(M. Schmitz, 2012) to search for relationships. Patterns and heuristics need to be calibrated therefore large corpus and significant seed samples are required. However, OIE systems (Fader, 2011),(Banko, 2007),(M. Schmitz, 2012) were created to efficiently parse large web resources and automatically detect all existing semantic relations. Efficient OIE system must possess following features (Niklaus, 2018):

- automation: relationships must be extracted without any prior training of them in unsupervised manner;
- corpus heterogeneity: relationships must be extracted in various genres of text;
- efficiency: extraction shall be possible in large web corpus.

Despite their success, current OIE systems are still faced with five main challenges pertaining to relationship extractions (Niklaus, 2018):

- overlaps: where same sentence fragment may generate multiple triples (subject, relation, object), for example, because of shared subject;
- discontinuation: where a text spans supporting the triples are is separated by an interval i.e. distant direct object;
- nesting: where one relation contains other relation as in complex sentences;
- distribution: where subject, relation and direct object mentions break a sentence and they are in different contexts. It is natural for a human to identify such cases, however, this still posses a challenge for current extraction and classification solutions;
- reliability: in general they do not return meaningful results on too small corpus

This paper describes a method to represent a short texts as a Semantic Frame Graph, which will:

- provide a method to limit negative impact of overlaps, discontinuation, nesting and especially distribution in relationship extraction;
- allow identifying relationships in short texts;
- fulfill automation and corpus heterogeneity requirement;

Gałusza, M.
A Semantic Frame Graph for Information Extraction.
DOI: 10.5220/0011729300003411
In *Proceedings of the 12th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2023)*, pages 435-442
ISBN: 978-989-758-626-2; ISSN: 2184-4313

- use Semantic Role Labeling to pre-process a text instead of complex dependency parsing;
- compare relation extraction efficiency using dialog coherency approach.

2 RELATED WORK

2.1 Open Information Extraction

The first solution formally introducing OIE was TextRunner (Banko, 2007). Its primary goal was to obviate manual adjustment of the extraction rules in case of corpora or a target relation shift. TextRunner was making a single pass over the corpus heuristically extracting relations triple centered around a verb phrase (VP) with a specific subject and direct object. It attempted to find the best arguments for that relation applying additional heuristics e.g. neither head nor target consist solely of a pronoun. A Naive Bayes was used as an estimator of a confidence function that was then trained over a set of features on the extractions so that the system could provide calibrated confidence values. Comparing to the last pre OIE solution, KnowitAll, quality of extraction and its performance significantly improved. Apart from that in KnwoItAll relations had to be specified upfront (Oren Etzioni, 2005). ReVerb (Fader, 2011) introduced additional syntactic and lexical constraints to limit incoherent and uninformative extractions of TextRunner in detecting distant relationships. For example: *"The Obama administration is offering only modest greenhouse gas reduction targets at the conference."* would yield a relation *"is offering only modest greenhouse gas reduction targets at"* between *"The Obama administration"* and *"the conference"*.

Fast dependency parsers and their ability to create a sentence Dependency Tree (DT) allowed construction of more sophisticated templates that further increased precision and recall of extraction. OLLIE (M. Schmitz, 2012) defined "Open Relation Patterns", which, using a dependency tree, were mediated by nouns and adjectives, not just verbs. OLLIE's processing began with seed tuples from REVERB and used them to build a bootstrap training set. It learned open pattern templates applied to individual sentences at extraction time.

The latest advancement in OIE is related to the latest progress in language modeling. The Transformer architecture leads to a novel paradigm of Neural Open Information Extraction (NOIE), (Zhou, 2022). NOIE approaches extractions from two major directions: Tagging and Generation.

Tagging-based solutions use annotation of tags' sequence corresponding to facts in the input sentence. The Generation-based ones directly decode relations relying on Sequence2Sequence architecture. Both tagging and generation paradigms predicts relationships auto-repressively, which means the current prediction relies on the previous output (Zhou, 2022). A skewed prediction will be inherited and magnified in the later steps. As the number of steps grows, errors accumulate and may decrease the performance.

An exemplary graph-based approach (Yu, 2021) breaks the auto-regressive factorization by constructing a graph where nodes are text spans and edges connecting them indicate that they belong to the same fact. Relationship discovery task is cast as maximal clique detection.

2.2 Semantic Role Labeling

Frame Semantics was originally introduced by Charles J. Fillmore (Charles, 1977) with the basic idea that one cannot understand the meaning of a single word without access to all the essential knowledge related to that word, namely, its semantic frame. The semantic frame is strictly associated with the word's meaning expressed in the sentence. Semantic Role Labelling (SRL) (D.Gildea, 2000) (D.Jurafsky, 2022) identifies and models frame's structure. SRL takes a sentence and identifies verbs and their arguments. Then, it classifies the arguments by mapping them to roles relevant to the verb in that frame, such as agent, patient, instrument, or benefactor. In other words, SRL tries to identify "Who, What, Where, When, With What, Why, How" for each frame. A state-of-the-art deep pre-trained SRL model (Peng Shi, 2019) detects the simplified structure of a frame where instead of an agent, a patient, an instrument it detects generic simplified arguments of a verb: ARG0, ARG1 and others. The structure of the frame highly correlates with the dependency tree (DT) of the sentence (T.Shi and O.Irsoy, 2020), where the verb and verb's arguments create constituencies (noun phrases NPs and verb phrases VPs). Moreover, it is possible to reduce the SRL task to a Dependency Parsing task (T.Shi and O.Irsoy, 2020) . In addition, SRL offers an efficient approach to the problem of the decomposition of complex sentences which was initially solved by a trained, dedicated, classifier splitting a sentence into shorter utterances (Angeli, 2019).

2.3 Knowledge Graphs

Constructing an ontology from text is challenging due complexity of human language. Initial approach

to construct a knowledge graph relied on syntactical parsing for terms, synonyms, concepts, relationships between them, their hierarchies. On top of them a set of axioms were created (or inferred from text) to be a set of logical implications constraining the interpretation of concepts and relations therefore governing inference (D. Maynard, 2017), (Cimiano, 2006).

A semantic frame serves two purposes: it is a means to abstract a cognitive schemata and it is this schemata computational counterpart (A. Gangemi, 2010). Structure of semantic frame naturally identifies objects in the sentence (as they are frame's arguments) and relationship between them (as it is a verb) (M. Alam, 2021). There are two limitations to that approach. First, neither concepts nor their hierarchies can be automatically detected. Second, connecting frames needs to be contextualized and supported by the text, meaning that connections mediated by co-occuring terms often require additional validation to make sure they form a logical flow. For example sentences: *"Astra Zeneca was first to develop a Covid19 vaccine. Covid19 was a serious threat to a global health in 2020 and 2021"* may generate a relationship between *Astra Zeneca* and *threat to a global health* through *Covid19*. Therefore a situation needs to be reconstructured from original text in order to validate the relationship (M. Alam, 2021). This paper addressed this problem specifically using dialog coherence approach to measure quality of connection between frames.

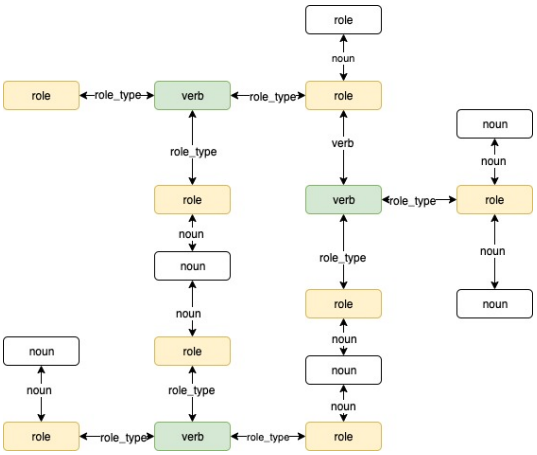

Figure 1: Semantic Frame Graph Structure.

A Semantic Frame Graph decomposes a sentence into a hierarchical structure of its frames.

SRL groups words per detected structure of the frame and yields a structure that correctly segments a sentence, solving, for example, the TextRunner's distant relationship issue (Figure 2). It does it without implementing any additional constraints and correctly detects that *"conference"* is a location, not a direct object of the verb: *"offer"* as in exemplary sentence: "The Obama administration is offering only modest greenhouse gas reduction targets at the conference." A corresponding SFG captures the structure detected by SRL (Figure 3).

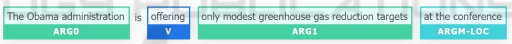

Figure 2: Example Semantic Decomposition.

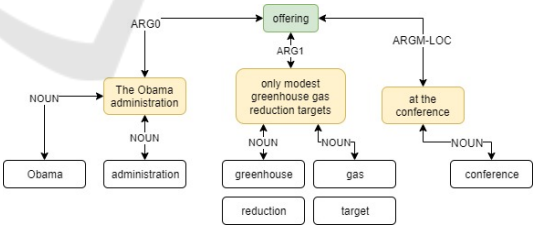

Figure 3: SFG Capturing Semantic Decomposition.

3 SEMANTIC FRAME GRAPH

Formally, a Semantic Frame Graph is an undirected, attributed, heterogeneous graph

$$G = (V, E)$$

where:

- V is a set of nodes of following types:
  - Noun : nouns detected in a frame.
  - Argument : a span of text describing semantic role of a frame
  - Frame : verb identifying a frame.
- E is a set of edges that represents a specific semantic role type: ARG0, ARG1, ...,. An edge 'VERB' is used if argument is further split into a frame; an 'NOUN' edge links nouns in the argument in case there is no more frames.

The graph is constructed by applying SRL identification (Peng Shi, 2019) on every sentence in the corpus. A general structure of the graph is depicted in Figure 1.

SFG decomposes complex sentences and in case of overlaps and discontinuations properly linking remote subjects and objects. Frames are always fully defined in its direct neighborhood which means that verbs without arguments are dropped. Comparing to the initial attempt (Angeli, 2019), where this task was cast as a linguistically driven search problem over a sentence DT, SFG relies in the SRL decomposition. For example, a sentence (Angeli, 2019): *"Born in a small town, she took the midnight train going anywhere"* is parsed to SFG (figure 4):

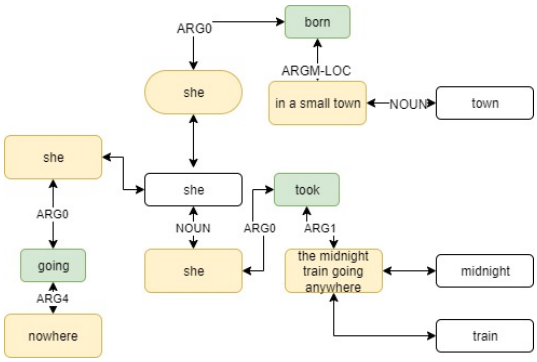

Figure 4: SFG Complex Sentence Decomposition.

The direct neighborhood of verbs creates complete frames, which are similar to a target shorter utterances (Angeli, 2019). It captures directly expressed relationships as: *"she born in a small town"*, *"she took the midnight train going anywhere"*, *"she going anywhere"*. Although the third utterance does make sense, it is incorrect from DT parsing perspective, I will discuss it in summary. Sentence decomposition correctly identifies *"she"* as a shared subject for the first and the second utterances and skipping *"took the midnight train"* completely for the third utterance which separates subject with the verb.

The SFG captures a hierarchy of frames as in a sentence *"A water landing of a jetliner that lost both engines due to hitting birds became known as the Miracle on the Hudson River"* (Figure 5)

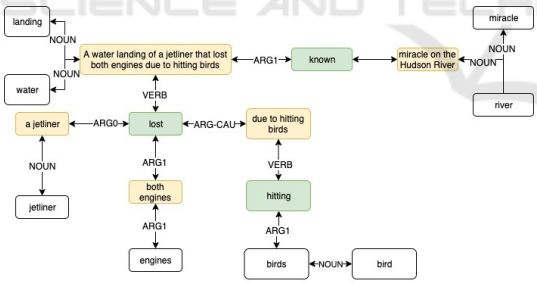

Figure 5: SFG Hierarchy of Frames.

SFG allows modeling distant relationships in the text that beyond single sentence. It allows linking mentions in the text that are outside of defined context window. The following example takes sentences describing a specific type of jet fuel. *"An airplane uses engines for flying. ATF is a type of aviation fuel designed for use in aircraft powered gas-turbine engines. (...) If these supercooled droplets collide with a surface they can freeze and may result in blocked fuel inlet pipes."*

The relationship between entities (nouns) exists if there is a path in SFG joining them. A path of semantic frames joining them will create a passage that will define a relationship that can be classified.

For example a path between water and an engine: *"water"* → *"supercooled water droplets collide with a surface"* → *"if supercooled water droplets collide with a surface they may result in blocked fuel inlet pipes"* → *"blocked fuel inlet pipes"* → *"aviation fuel designed for use in aircraft powered by gas turbine engines"* → *"an airplane uses engines for flying"* → *"engine"*

The generated path may indicate that *"water"* impacts *"engine"* that may also impact the safety of *"flight"*

4 SHORT TEXT REPRESENTATION

In this section will compare an SFG graph with a Sentence Graph (SG) which does not perform decomposition. I will use a short text (12 sentences) describing news excerpt on Iraq War as in (D.Radev, 2004). The SG approach (D.Radev, 2004) constructs a weighted graph of sentences where connections are defined by co-occurring nouns and connection weight equals the similarity between two sentences defined by idf-modified-cosine. The threshold on the similarity manages connectivity of the graphs and hence quality of the paths.

A Table (1) summarizes structure. A sample context, around noun "Baghdad" is provided in Figures (7) and (8). An SFG represents semantics of the text in more compact manner and even without any thresholds it can reduce number of paths almost 76 times. In general, the problem of detecting a relationship between entities can be cast as finding a path between them. Structure of the graph impacts the performance of relationship extraction.

Table 1: Graphs' structure comparison.

metric	SG	SFG
number of nodes	144	194
number of edges	193	219
number of noun nodes	76	76
number of verb nodes	57	60
number of paths	**3901764**	**51040**

A centrality (PageRank) of words (Table 2) shows that 8 out of 10 top words are preserved in an SFG representation so there is insignificant semantic drift between both representations.

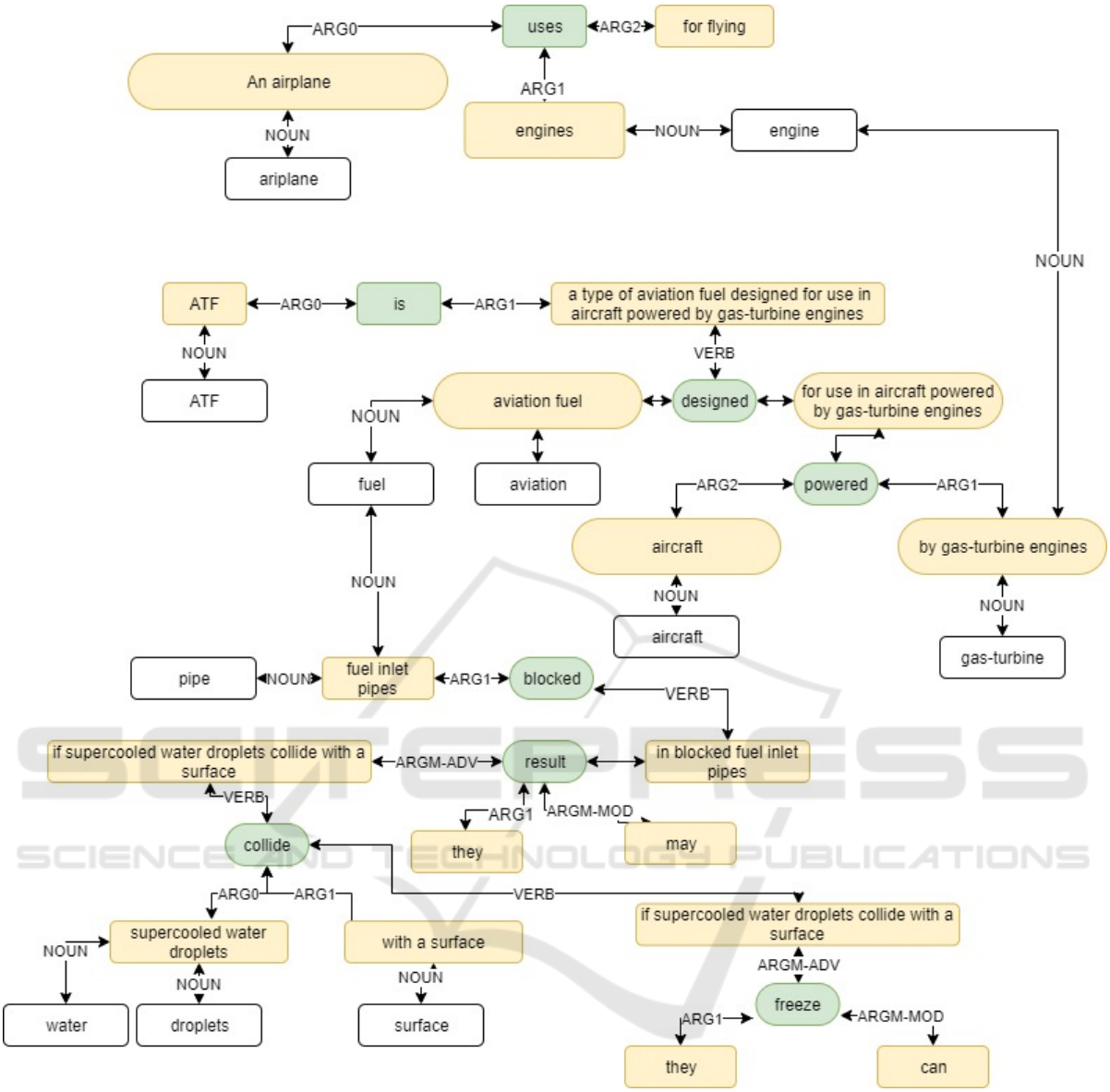

Figure 6: SFG Distant Relationships.

4.1 Coherence of Paths and Relationship Extraction

SFG represents frames which are singular sentences. Therefore a path between any entities is a sequence of generated singular sentences. Such a sequence can be measured for its coherence.

The SFG represents a set of N frames

$$F : \{f_1, f_2, ..., f_N\}$$

if the similarity of frames has the property that

$$s : \forall f_i, f_j \in F, 0 \leq s(f_i, f_j) \leq 1 \quad (1)$$

then the coherence c of a path p containing K frames I define as the minimum similarity of the neighboring frames in the path:

$$c_p : \min_{0 \leq j < K} s(f_j, f_{j+1}) \quad (2)$$

There are various realizations of function (1) available: sentence similarity (N. Reimers, 2019), textual entailment (Poliak, 2020).

I update (B. Grosz, 1995) approach to meet requirements of function (1) and I replace a direct noun expressions to evaluate continuation retaining and shifting with a semantic approach.

I use the fact that frame's subject and object (centers) are decomposed into ARG0 and ARG1 roles

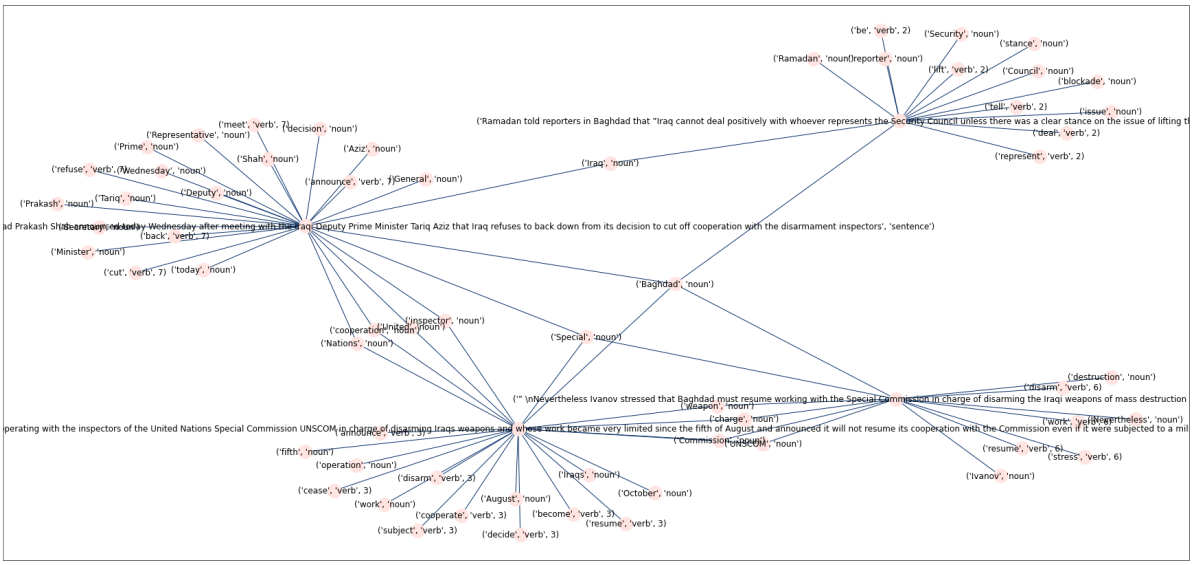

Figure 7: A "Baghdad" SG Context.

Table 2: Top 10 key words comparison.

pageranked position	sentence	frame
1	Iraq	Iraq
2	United	United
3	Nations	today
4	today	Baghdad
5	Prime	Prime
6	Baghdad	inspector
7	Minister	Minister
8	inspector	Nations
9	Blair	year
10	Ivanov	work

and I update the original centering approach using zero-shot label classification. Instead of continuation/retention/shift signals I evaluate cosine distance between classification of backward centers of current frame vs. previous.

The Wikipedia's page for Teton Dam http://en.wikipedia.org/wiki/Teton_Dam is used as an example to identity if selected object (*a storage tank*) posed a risk and what kind of risk to another object (*a city*) in the event of Teton Dam collapse.

Percentage of coherent path give minimal threshold (Figure 9) shows that RTE immediate drop indicating that frames are rather not each other entailments and coherent passage will have low coherence score making it difficult to find.

Best paths with best coherence score for each evaluation method (tab. 3) show that modified centering approach on the SFG linked all related sentences from original text to form a passage confirming that there is a semantic relationship between *a storage tank* and *a city*.

Table 3: Best paths for coherence types.

path	c_p	type
Dozens of logs hit a bulk gasoline storage tank a few hundred yards away → When the flood waters hit thousands of logs washed into town → When the flood waters hit → The city of Idaho Falls even further down on the flood plain had time to prepare	0.52	semantic similarity
Dozens of logs hit a bulk gasoline storage tank a few hundred yards away → The gasoline sent flaming slicks adrift on the racing water → When the flood waters hit → The city of Idaho Falls even further down on the flood plain had time to prepare	0.21	RTE
Dozens of logs hit a bulk gasoline storage tank a few hundred yards away → The gasoline sent flaming slicks adrift on the racing water → When the flood waters hit thousands of logs washed into town → The force of the logs and cut lumber and the subsequent fires practically destroyed the city	0.85	centering

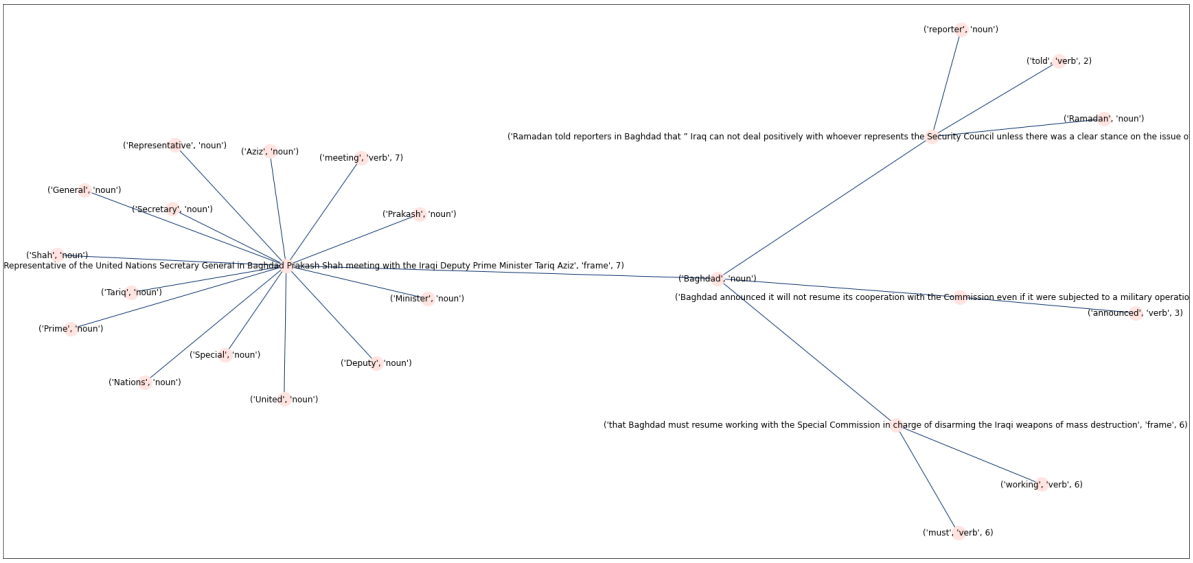

Figure 8: A ”Baghdad” SFG Context.

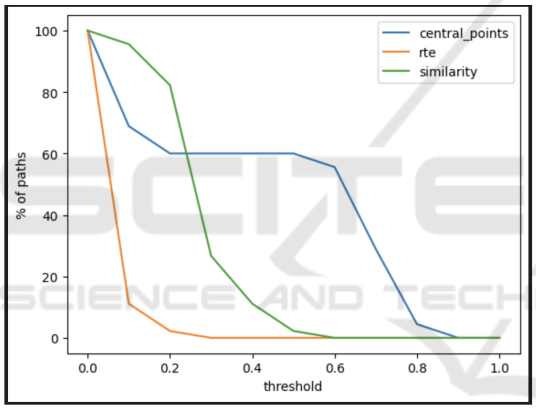

Figure 9: Fraction of coherent paths given threshold.

5 CONCLUSION

A Semantic Frame Graph is an alternative representation of text. It uses SRL as pre-processing step to identify the structure of frames. A whole set of frames is loaded into a graph that is used as a foundation for identifying relationships between selected words. It links mentions distant in text, even across sentences. It does not use a complex rule-based approach that requires bootstrapping nor significant corpus to validate them as in early OIE solutions. Even without any additional edges' weighting and thresholds on them, it shows significant reduction in number of paths between entities so further classification of them to proper relationships will require less computation. Modified centering approach measures overall coherence of paths between entities and thresholds enables control of the quality of the connection that can be classified for its type afterwards.

A few issues can be solved to increase text representation quality further. Current state-of-the-art SRL parser (Peng Shi, 2019) can be augmented with additional post-processing steps to correct the alignment of SRL tags with a sentence's dependency tree. In exemplary sentence (Figure 5), for detected frame ”she going nowhere” a subject is *”train”* which is visible in its dependency tree. A progress in a coreference resolution will further increase the quality of the graph not only because pronouns will be properly replaced by nouns they refer to, but also descriptive expressions like *”this event”*, *”in this case”* will properly link the frame with a noun or noun phrase mention in the text.

REFERENCES

A. Gangemi, V. P. (2010). Towards a pattern science for the semantic web. In *Semantic Web Journal*. IOS Press.

Angeli, G. (2019). Leveraging linguistic structure for open domain information extraction. In *Proceedings of the Association of Computational Linguistics*. ACL.

B. Grosz, A. Joshi, S. W. (1995). Centering a framework for modeling the local coherence of discourse. In *Computational Linguistics*. MIT Press.

Banko, M. (2007). Textrunner: Open information extraction on the web. In *ACL-HLT*.

Charles, F. (1977). Scenes-and-frames semantics. In *In A. Zampolli, ed. Linguistic Structures Processing*.

Cimiano, P. (2006). Ontology learning and population from text algorithms, evaluation and applications. Springer.

D. Maynard, K. Bontcheva, I. A. (2017). Natural language processing for the semantic web. In *SYNTHESIS LECTURES ON THE SEMANTIC WEB: THEORY AND TECHNOLOGY #15*. Morgan & Claypool.

D.Gildea, D. (2000). Automatic labeling of semantic roles. In *Association for Computational Linguistics*. ACL.

D.Jurafsky, J. (2022). *Speech and Language Processing*. DRAFT, Stanford, 2nd edition.

D.Radev, G. (2004). Lexrank: Graph-based lexical centrality as salience in text summarization. In *Journal of Artificial Intelligence Research*. ACL.

Fader, Soderland, E. (2011). Identifying relations for open information extraction. In *ACL*.

Ji, S. (2021). A survey on knowledge graphs: Representation, acquisition and applications. In *IEEE Transactions on Neural Networks and Learning Systems*. IEEE.

M. Alam, A. G. (2021). Semantic role labeling for knowledge graph extraction from text. In *Progress in Artificial Intelligence*.

M. Schmitz, S. S. (2012). Open language learning for information extraction. In *ACL*.

N. Reimers, I. G. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. In *EMNLP*. EMNLP.

Niklaus, C. (2018). A survey on open information extraction. In *27th International Conference on Computational Linguistics*. ACL.

Oren Etzioni, A. Y. (2005). Unsupervised named-entity extraction from the web: An experimental study. In *Artificial Intelligence*.

Peng Shi, J. J. L. (2019). Simple bert models for relation extraction and semantic role labeling. In *ArXiv*. ALLENNLP.

Poliak, A. (2020). A survey on recognizing textual entailment as an nlp evaluation. In *EMNLP*. EMNLP.

T.Shi, I. and O.Irsoy (2020). Semantic role labeling as syntactic dependency parsing. In *ACL*.

Yu, B. (2021). Maximal clique based non-autoregressive open information extraction. In *EMNLP*.

Zhou, S. (2022). Survey on neural open information extraction: Current status and future directions. In *IJCA*.