Extracting Orientation Relations between Geo-Political Entities from their Wikipedia Text

Nitin Ramrakhiyani[1,2,*], Vasudeva Varma[1] and Girish Keshav Palshikar[2]

[1]International Institute of Information Technology (IIIT), Hyderabad, India
[2]TCS Research, Pune, India

Abstract

Augmenting Wikidata with spatial relations specific to Geography can be useful for increasing its utility in multiple applications. In this paper, we aim to extract orientation of borders between countries in the world, from their Wikipedia text and suggest its use to augment the *shares_borders_with* relation in Wikidata. We propose the use of Natural Language Inference (NLI) for extracting the orientation relations from text and show that when combined with contextual lexical patterns, the performance becomes better than the standard NLI setting.

Keywords

spatial information extraction, orientation relation extraction, zero-shot natural language inference, wikidata augmentation

1. Introduction

Spatial information about geo-political entities such as countries, states and counties, finds mention in the first few sentences of their Wikipedia page, stressing its importance in an entity's description. This spatial information can consist of topological (`country located in a continent`), orientation (`country bordered in a direction to another country`) and distal (`country having a certain area`) facts (Examples in Table 1). The structured counterpart of Wikipedia - the Wikidata knowledge base, also captures these information pieces through properties such as *shares_borders_with*, *located_in_the_administrative_territorial_entity* and *basin_country*, indicating the relation between entities and corresponding values (Table 1).

The *shares_borders_with* property is an important property which captures if an entity shares a land or a maritime border with another. An important aspect of the property is the orientation of the shared border with respect to the described entity. For example, `north` in (Denmark is ... lying ... north of Germany.) (Row 2 in Table 1). This information about border orientation is captured using the *direction_relative_to_location* qualifier of the *shares_borders_with* property. However, this is also one of the least captured aspects of this property and is only available for only about 10% of *shares_borders_with* instances between countries. As the first contribution of

GeoExt 2023: First International Workshop on Geographic Information Extraction from Texts at ECIR 2023, April 2, 2023, Dublin, Ireland

*Corresponding author.
nitin.ramrakhiyani@research.iiit.ac.in (N. Ramrakhiyani); vv@iiit.ac.in (V. Varma); gk.palshikar@tcs.com (G. K. Palshikar)
https://nramrakhiyani.wordpress.com (N. Ramrakhiyani)

Table 1
Examples of Spatial Wikipedia descriptions and Wikidata relations (for Denmark)

Spatial Information from Wikipedia description	Corresponding Wikidata relations
Denmark is a Nordic country in Northern Europe.	(Denmark, P30:*continent*, Europe)
Metropolitan Denmark is the southernmost of the Scandinavian countries, lying south-west of Sweden, south of Norway, and north of Germany.	(Denmark, P47:*shares_borders_with*, Sweden), (Denmark, P47:*shares_borders_with*, Norway), (Denmark, P47:*shares_borders_with*, Germany)
Spanning a total area of 42,943 km2 (16,580 sq mi),[9] metropolitan Denmark consists of the northern part of ...	(Denmark, P2046:*area*, 42,925.46±0.01 square kilometre)

the paper, we aim to fill this important gap specifically for borders between countries, through automatic extraction of orientation information from their Wikipedia text.

Multiple benchmarks [1, 2, 3] and approaches [4, 5, 6, 7] for extracting spatial information from text have been proposed. Most of the approaches are primarily supervised and use the sequence labelling paradigm employing Conditional Random Fields. Recent advances in NLP have been driven by the Transformer based Large Language Models (LLMs). We believe that LLMs can be harnessed for extraction of such spatial information and more importantly, in an unsupervised setting. As the second contribution, we propose the use of Natural Language Inference (NLI) based information extraction carried out on LLMs for identifying orientation relations from spatial text about geographical entities (currently limited to countries in this paper). We also boost the NLI based approach by enhancing the hypothesis templates through patterns which capture the lexical context of the relation. We present a comparison with a few relevant baselines such as zero shot prompt tuning [8] and prompt tuning with demonstrations [9] and show that the NLI approach boosted with lexical patterns is the most promising.

2. Motivation and Problem Description

Augmenting the Wikidata KB with missing entities, relations and values of relation qualifiers is useful for (i) increasing accessibility of structured knowledge facts both for humans and automatic agents, and (ii) generation of textual content (say a Wikipedia page) for rare entities/relations. Moreover, if the augmentation can be automated and based on robust techniques, the KB completion can be faster and accurate. The current work is a step in this direction, involving augmentation of Geography knowledge to the Wikidata KB, through an automatic unsupervised text mining approach.

The overall problem is to supplement the *shares_borders_with* property between countries with a qualifier *direction_relative_to_location* wherever it is not added to the property. The qualifier, particular to this property, describes the orientation of the border with respect to the main entity. For example, in the Wikidata entry for France[1], it is shown to share land borders with 10 countries, but the qualifier is available only for one case ('north' with Belgium). However, in the Wikipedia page for France, the sentence - `Its land borders consist of Belgium and Luxembourg in the northeast, Germany and Switzerland in the east, Italy`

[1]https://www.wikidata.org/wiki/Q142

`and Monaco in the southeast, and Andorra and Spain in the south and southwest,` clearly indicates the other border orientations, not captured in France's Wikidata KB relations.

At a finer level, the problem, given the above sentence, is to extract the direction relations between the described subject entity, (also known as *Trajector* in Spatial Information Extraction literature) with other entities (*Landmark*) mentioned in the sentence. So in this example, the goal is to extract relations such as [Its (France), northest, Luxembourg], [Its (France), east, Germany] and [Its (France), south, Spain]. Once these relations are extracted, the augmentation of the corresponding *shares_borders_with* properties for France can be supplemented with the correct *direction_relative_to_location* qualifiers. This process can be carried out similarly for all countries and their corresponding property

3. Proposed Approach

The standard approaches in Spatial Information Extraction are primarily supervised and work for specific tasks. One of our earlier work [6], is a supervised approach which uses a two step neural network, one for entity extraction and another for relation extraction. However, it has a limited focus of extracting spatial relations from image captions which are simple sentences and not complex as the one above (describing France's borders). A more recent approach such as Shin et al. [7] also uses a similar two step approach based on BERT representations. However, apart from being a supervised approach, it involves training data from the SemEval-2015 task. The data is more complex but not directly relevant for extracting orientation relations in Geography texts. Another recent approach by Wang et al. [10] also has similar constraints apart from being a complex technique[2]. We believe that these approaches are closed with respect to the problem they solve and work with a different kind of training data. Though these techniques use word embeddings and even PLM based representations (such as BERT), they do not harness the knowledge captured in these PLMs explicitly, which can be beneficial for the task of extracting Geography knowledge. Apart from being sources of knowledge, PLMs support multiple unsupervised approaches for extraction of information, through probing tasks. We propose and explain the use of PLMs for extracting the orientation relations in a zero-shot setting, through a Natural Language Inference (NLI) approach.

3.1. Zero-shot NLI based Relation Extraction

PLM based NLI requires the text under consideration to be posed as a *premise* which is coupled with a suitable *hypothesis* for a textual entailment task. The task involves classifying the premise-hypothesis pair into whether the hypothesis logically follows from the premise (*Entailment*) or contradicts it (*Contradiction*) or isn't related to it (*Neutral*). The hypothesis allows the inclusion of the class information in the premise-hypothesis pair and hence, devising the hypothesis becomes an important part of the exercise. A hypothesis comprises of two customizable parts - a template and a class-indicating phrase. In this paper, we use the direction names as the class-indicating phrases. The template can be devised using two approaches.

[2]Code for both these recent approaches is unavailable and hence both require a separate significant reproducibility effort, which has been kept as later work.

3.1.1. Using a basic template

As a basic hypothesis template, we propose the use of the arrangement "*Trajector* shares borders with *Landmark* to the *Direction*". So given the input text `Denmark is lying southwest of Sweden, south of Norway, and north of Germany.` and one of the directions/classes (say north), the premise-hypothesis pair will be (P: `Denmark is lying southwest of Sweden, south of Norway, and north of Germany.`, H: `Denmark shares borders with Germany to the north.`). Such pairs can be created for all directions and hence, for each Trajector and Landmark entity pair, 8 such premise-hypothesis pairs will get created.

3.1.2. Templates using mentions of the property

A detailed observation of different sentences reveals that there is a specific way in which the borders relation is mentioned in a sentence. For example, in the Denmark example, `... is lying to the north of ...` is used to express the borders relation, which is different from Afghanistan, where `... is bordered by ... to the east` is used. We identified about 10 different higher level mention styles, a subset of which is shown in Table 2 in the form of lexical patterns. To aid the PLM in performing an informed entailment, we propose to adapt the hypothesis template based on the pattern used in the premise. So in this example, the hypothesis for Denmark will be constructed using the "lying" pattern and for Afghanistan will use the "bordered by" pattern. We hypothesize that combining the power of classical lexical patterns with PLM's attention mechanism and pre-training can lead to better results than using basic/standard hypothesis templates as above.

In the zero-shot setting, a PLM pretrained for the NLI task, is then fed each of these instantiated premise-hypothesis pairs and the PLM predicts whether there is an *Entailment*, *Contradiction* or *Neutral* between them. The premise-hypothesis instance for which the PLM predicts an Entailment with the highest confidence is selected and the hypothesis' corresponding direction is predicted as the final class for the premise/input text.

Table 2
A subset of the identified patterns through which the borders relation is mentioned

Trajector	shares land borders with	*Landmark*	to the	*Direction*
Trajector	is bordered to the	*Direction*	by	*Landmark*
Trajector	is bordered by	*Landmark*	to the	*Direction*
Trajector	bounds	*Landmark*	to its	*Direction*
Trajector	is lying to the	*Direction*	of	*Landmark*
Trajector	is located	*Direction*	of	*Landmark*

4. Experimentation and Evaluation

4.1. Baselines

As the baseline, we consider a prompt based mask filling approach. We use the hypothesis sentences, created as part of the proposed approach (Section 3.1.2), as the prompts where in

place of the *Direction* we keep the [MASK] token. We employ PLMs which are trained for a Masked Language Modelling (MLM) task, and get them to fill the correct direction in place of the MASK token. To provide the premise-like support, we devise another baseline on lines similar to the prompting with demonstrations idea proposed in Gao et al. [9]. In this case we build the prompt by concatenating the premise to the MASK sentence with a [SEP] token, thereby indirectly demonstrating it with what needs to be filled in the mask. We hypothesize that this should give the PLM the necessary context for filling the [MASK] token.

4.2. Dataset

As the first step in the experimentation, we construct a dataset of sentences which emit direction information. Based on these sentences we create the premise hypothesis pairs to be consumed by the NLI approach and the masked prompts for the baseline.

- We first use a SPARQL query on Wikidata to find the list of countries which have a *shares_borders_with* property. This led to creation of a list with 177 countries.
- We use the MediaWiki API[3] to obtain the first 10 sentences of the Wikipedia page of each of the countries in the list. In all, the total number of sentences obtained were 1766.
- We then devise a simple rule to retain any sentences which have atleast one direction and two country names present and filter out the rest of the sentences from each entity's sentences. We further filtered sentences, which didn't communicate a bordering relation. After this filtering, the total number of sentences left was 143.
- It is important to note that, this filtering leaves us with sentences which refer to the Trajector/main entity with pronouns such as `It` or common nouns such as `The country`, `The nation` and `The archipelago`. For creating proper premises and prompts, we replaced these mentions in the sentences with the proper Trajector entity names.

For the creation of the gold standard dataset, all bordering relations in the 143 sentences were labelled. Each gold relation is of the form (Trajector, Direction, Landmark) and to be read is *The Trajector entity* shares border with *the Landmark entity* to the *Direction*. So for example, a gold relation (Denmark, south, Germany) is understood as Denmark shares border with Germany to the south. A total of 562 gold relations were labelled in this manner.

4.3. Experimentation

For the NLI approaches, we experiment with the bart-large-mnli[4] and roberta-large-mnli[5] models from the huggingface library. Both these models are standard transformer models, tuned further on the Multi-NLI dataset [11] through which they gain the necessary NLI capability. For the prompting baselines, we use the BERT_{Large} [12] and BART_{Large} [13] models. The gold standard data and code can be obtained from the corresponding author through an email request.

[3] https://en.wikipedia.org/w/api.php
[4] https://huggingface.co/facebook/bart-large-mnli
[5] https://huggingface.co/roberta-large-mnli

4.4. Evaluation

In Table 3, we report the results for the NLI approach (basic and patterns template) and the prompting baselines, in terms of Precision@k (k = 1 and 3) which checks if the correct direction is predicted at the k^{th} rank in the predictions.

Table 3
Comparative Results

	P@1	P@3	P@1	P@3
	bart-large		berta-large-cased	
Zero-shot prompting	0.1563	0.4463	0.2034	0.4972
Zero-shot prompting (with demo)	0.1130	0.4840	0.0697	0.2976
	bart-large-mnli		roberta-large-mnli	
NLI (basic template)	0.8329	0.8738	0.8399	0.8757
NLI (pattern template)	0.8964	0.9190	**0.9096**	**0.9171**

As can be observed in Table 3, the NLI approach is significantly better than the prompting baselines irrespective of the models. Also as hypothesized, adapting the NLI hypotheses with lexical patterns helps in improving the performance by 4 to 8%. With respect to the NLI PLMs, the roberta-large-mnli model performs slightly better than the bart-large-mnli model.

Some peculiar difficulties observed in the task are: (i) Some countries are referred to differently in different contexts and hence the hypothesis creation is missed for non-listed mentions. For example, the Wikidata entry titled *People's Republic of China* is the official reference of China. However, it is referred to simply as *China* in descriptions of its neighbours such as India. (ii) Some countries have names which are subsets of other countries and hence, unwanted mentions of those countries get considered during hypothesis creation. For example, *The Democratic Republic of Congo* and the *Republic of Congo*; or *Sudan* and *South Sudan*. We also investigated why prompting based approaches are performing so poorly, in spite of already devised using the lexical patterns. Firstly, the token predicted at the MASK token is many times not even a direction mention. For example, `Denmark is lying [MASK = within] of Norway`. Secondly, single word directions such as north, east are predicted more than the bi-word directions such as northeast, leading to incorrect predictions for such bi-word orientations.

5. Conclusion

We aim at adding orientation information to Wikidata KB's *shares_borders_with* property between countries of the world. We demonstrated the use of Natural Language Inference (NLI) based querying of PLMs for this relation extraction task. We also showed that adapting the NLI hypotheses based on the mention of the property in the input text, boosts the performance further. Overall, we envisage that using NLI based techniques can be a promising direction for spatial information extraction, particularly relating to geographical and geo-political entities.

References

[1] P. Kordjamshidi, S. Bethard, M.-F. Moens, SemEval-2012 Task 3: Spatial Role Labeling, in: Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, Association for Computational Linguistics, 2012, pp. 365–373.

[2] O. Kolomiyets, P. Kordjamshidi, M.-F. Moens, S. Bethard, Semeval-2013 Task 3: Spatial Role Labeling, in: Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2, 2013, pp. 255–262.

[3] J. Pustejovsky, P. Kordjamshidi, M.-F. Moens, A. Levine, S. Dworman, Z. Yocum, SemEval-2015 Task 8: SpaceEval, in: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), 2015, pp. 884–894.

[4] P. Kordjamshidi, M. Van Otterlo, M.-F. Moens, Spatial Role Labeling: Towards Extraction of Spatial Relations from Natural Language, ACM Transactions on Speech and Language Processing (TSLP) 8 (2011) 4.

[5] A. Mazalov, B. Martins, D. Matos, Spatial Role Labeling with Convolutional Neural Networks, in: Proceedings of the 9th Workshop on Geographic Information Retrieval, ACM, 2015.

[6] N. Ramrakhiyani, G. Palshikar, V. Varma, A Simple Neural Approach to Spatial Role Labelling, in: European Conference on Information Retrieval, Springer, 2019, pp. 102–108.

[7] H. J. Shin, J. Y. Park, D. B. Yuk, J. S. Lee, BERT-based Spatial Information Extraction, in: Proceedings of the Third International Workshop on Spatial Language Understanding, 2020, pp. 10–17.

[8] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, G. Neubig, Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing, ACM Computing Surveys 55 (2023) 1–35.

[9] T. Gao, A. Fisch, D. Chen, Making Pre-trained Language Models better Few-shot Learners, arXiv preprint arXiv:2012.15723 (2020).

[10] F. Wang, P. Li, Q. Zhu, A Hybrid Model of Classification and Generation for Spatial Relation Extraction, in: Proceedings of the 29th International Conference on Computational Linguistics, 2022, pp. 1915–1924.

[11] A. Williams, N. Nangia, S. Bowman, 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), Association for Computational Linguistics, 2018, pp. 1112–1122. URL: http://aclweb.org/anthology/N18-1101.

[12] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, arXiv preprint arXiv:1810.04805 (2018).

[13] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension, arXiv preprint arXiv:1910.13461 (2019).