Entity Linking in the ParlaMint Corpus

van Heusden, R.; Marx, M.; Kamps, J.

Publication date
2022
Document Version
Final published version
Published in
ParlaCLARIN III : Workshop on Creating, Enriching and Using Parliamentary Corpora : proceedings

Proceedings of ParlaCLARIN III @ LREC2022 pages 47–55
Marseille, 20 June 2022

Entity Linking in the ParlaMint Corpus

Ruben van Heusden, Maarten Marx, Jaap Kamps
University of Amsterdam
Science Park 904, 1098XH
r.j.vanheusden@uva.nl, maartenmarx@uva.nl, kamps@uva.nl

Abstract

The ParlaMint corpus is a multilingual corpus consisting of the parliamentary debates of seventeen European countries over a span of roughly five years. The automatically annotated versions of these corpora provide us with a wealth of linguistic information, including Named Entities. In order to further increase the research opportunities that can be created with this corpus, the linking of Named Entities to a knowledge base is a crucial step. If this can be done successfully and accurately, a lot of additional information can be gathered from the entities, such as political stance and party affiliation, not only within countries but also between the parliaments of different countries. However, due to the nature of the ParlaMint dataset, this entity linking task is challenging. In this paper, we investigate the task of linking entities from ParlaMint in different languages to a knowledge base, and evaluating the performance of three entity linking methods. We will be using DBPedia spotlight, WikiData and YAGO as the entity linking tools, and evaluate them on local politicians from several countries. We discuss two problems that arise with the entity linking in the ParlaMint corpus, namely inflection, and *aliasing* or the existence of name variants in text. This paper provides a first baseline on entity linking performance on multiple multilingual parliamentary debates, describes the problems that occur when attempting to link entities in ParlaMint, and makes a first attempt at tackling the aforementioned problems with existing methods.

Keywords: entity linking, multilingual, ParlaMint

1. Introduction

The ParlaMint corpus was created by CLARIN[1] in order to facilitate multilingual research on parliamentary proceedings, with the original project concerning four countries, which was later increased to seventeen countries and counting (Erjavec et al., 2021). The goal of the ParlaMint project is the unification of parliamentary debates across European countries, facilitating research of these documents by researchers across various disciplines. The ParlaMint subcorpora consist of both 'plain text' and annotated versions, with the annotated versions containing automatically annotated Part-of-Speech tags, lemmas, Named Entities, as well as a variety of other linguistic features. These Named Entities can be of particular interest to researchers, as they provide them with a landscape of actors and objects present in the dataset, as well as the relationships between these entities.

Although a wide variety of entity linkers is available today, the case of linking Named Entities in the ParlaMint corpus to an existing knowledge base is of a different nature than most other Entity Linking (EL) tasks. Not only are the entities in four different alphabets, some languages lack solid coverage by the EL systems, and many countries have different morphologies and are rich in inflections. Moreover, we are dealing with real world data, and as such some of the entities might be misspelled or ambiguous, or strings that are not a Named Entity are mistakenly tagged as Named Entity. Such mistakes mostly consist of strings being tagged that are too generic, for example 'Mr Speaker', complicating the linking process, or entities being tagged with the incorrect entity type. Although the parliamentary proceedings of the countries in ParlaMint are carefully curated, spelling mistakes do occur on rare occasions. For example in the Dutch subcorpus, several names containing the 'ö' character are written with 'oe' instead, or vice versa, or the name 'pechtold' is reported as 'pechtol'. This is amplified by a problem that is quite specific to spoken text and by extension parliamentary debates, best described as *aliasing* or the existence of name variants. The problem of aliasing occurs when actors are not mentioned with their full name, but for example only their surname, or a nickname. For example 'Joe Biden' might be referred to as 'Mr. Biden', which complicates the linking process, as not having the first name to work with significantly increases ambiguity.

In this research, we evaluate three existing Entity Linking systems, namely **DBPedia-spotlight**, **WikiData** and **YAGO** on the ParlaMint dataset, investigating the aforementioned problems.

Our research questions are as follows:

- **How well do three existing Entity Linking systems (DBPedia, WikiData, YAGO) work on parliamentary actors, such as those present in ParlaMint?** For this research question we extracted members of local parliaments from WikiData and extracted the unique *Q-item* identifiers to obtain gold standard data for individual countries, and provide a fair comparison across countries by having names without inflections and possible spelling errors. We evaluate the accuracy

[1] https://www.clarin.eu/content/parlamint-towards-comparable-parliamentary-corpora

of all three systems on the dataset and report the main differences between the three systems, while also focusing on the difference in performance of the linkers across languages. Hereby we extend on the works of Pillai et al. (2019) and Färber et al. (2015) by analysing the Entity Linking component of these three knowledge bases.

- **How much does lemmatization help with improving the performance on languages with a high number of inflections?** For languages such as Polish, Named Entities are inflected quite often, making the Entity Linking process more difficult. In this research question, we make use of the provided lemmas of the Entities in ParlaMint to investigate whether lemmatization can help improve the performance of Entity Linking systems, and when lemmatization is less effective.

- **How can the phenomenon of aliasing be counteracted?** One of the peculiarities of the ParlaMint dataset is the phenomenon of *aliasing*, where names are either abbreviated or nicknames are used. For example in the case of 'Joe Biden' and 'President Biden' or 'Biden'. In this research question we investigate two simple methods of counteracting the phenomenon. The first method works by searching for variants of the name at various levels, for example in debates in the same week, or debates in the same month. The other method uses the speaker metadata present in the ParlaMint corpora to match entities with members of parliament and other speakers.

2. Related Work

Regarding the case of Entity Linking in multiple languages, there have been several papers that address this issue (De Cao et al., 2021; Sil et al., 2018; Botha et al., 2020; McNamee et al., 2011; Pappu et al., 2017). Sil et al. (2018) introduce a neural method for performing entity linking in multiple languages. Their approach is to link entities from different languages to their corresponding entities in the English version of Wikipedia. To achieve this, they train a neural network that makes use of multilingual word embeddings to compare the contexts of entities and candidates, as well as using features such as the number of overlapping words. The model is trained on English entities, and tested on different languages to see how well this zero-shot setting works for the entity linking case. In their work they found that the model is able to achieve state-of-the-art performance on both the monolingual case and the multilingual case, given that multilingual embeddings are available for those countries.

De Cao et al. (2021) makes a clear distinction between the tasks of *Cross Lingual Entity Linking* (XEL) and *Multilingual Entity Linking*(MEL). In the case of Cross Lingual Entity Linking, candidates from different languages are all mapped to entities in a monolingual knowledge base. In the case of Multilingual Entity Linking, candidates from different languages are mapped into a multilingual knowledge base. In their paper they describe their MEL system, which consists of an auto-regressive seq2seq model for computing the context similarity between the entities in text and the candidate entities in the knowledge base.

Our work attempts to bypass the issues associated with language specific knowledge bases, by using the Q-items provided by WikiData as the means of verifying links from entities to the knowledge base. By using these items, entities in a specific language could be linked to the Q-ID of that item, even if the item is represented in another language.

The problem of aliasing, or the usage of name variants and partial names for different types of entities has been studied previously. One study that addressed this problem is Gottipati and Jiang (2011), which attempts to tackle, among other things, the problem of name variants. This is done by query expansion, where both knowledge from the query itself and external knowledge are used to resolve entities. To resolve entities using local knowledge, other named entities in the same document as the query entity are checked to see whether they contain the query entity as a substring. If so, then this entity is added to the query as an alternative variant. The algorithm used in our paper is quite similar to the method for adding local knowledge used in Gottipati and Jiang (2011), with the exception that our method is not limited to one document, but rather includes multiple documents based on the time window.

There have been a multitude of papers that compare different knowledge bases on various aspects, such as consistency and timeliness of information. (Färber et al., 2015; Pillai et al., 2019). Färber et al. (2015) compare several knowledge bases including WikiData, DBPedia and YAGO on a variety of aspects. These aspects include the number of languages included in the knowledge base, which domains are covered, the number of relations in the knowledge base and the whether or not correctness constrains are enforced in the knowledge base, among other criteria. They found that there are various differences between knowledge bases, mostly regarding the amount of information present for facts (such as a description or a source of the fact), but argue that the exact requirements needed for a knowledge base can vary depending on the specific task it is being used for.

3. Method

3.1. Q-Items

Q-items or Q-IDs are the identifiers used in WikiData for identifying unique entities and concepts in the WikiData knowledge base.[2]. These identifiers are cross

[2] https://www.wikidata.org/wiki/Wikidata:Glossary

lingual, meaning that for example 'Angela Merkel' will have the same Q-ID, whether the entity is searched in the English or German WikiData. Besides entities, Q-IDs are also given to attributes or properties of entities. For example 'Member of the European Parliament', which has Q-ID $Q27169$. These Q-IDs thus allow for the comparison of entities in different languages and different knowledge bases, given that the knowledge base in question also reports Q Numbers. For both DBPedia and YAGO this is true at least up to a degree, and for entities that do not have this Q-ID, the Q-ID can often be discovered through a WikiPedia link present for the entity.

3.2. Systems

We evaluate three Knowledge Bases / Entity Linking systems: **DBPedia**, **WikiData** and **YAGO**. Below we describe them briefly.

3.2.1. DBPedia

The API from *DBPedia spotlight* (Mendes et al., 2011) is used to detect and link entities in text to the DBPedia knowledge base. In the API the 'candidates' call is used to retrieve candidates for the entity, and the default parameters are used. To link entities from DBPedia with WikiData, we retrieve the Q-items from the entities in DBPedia using the $< owl : SameAs >$ property. If the entity does not have a Q-item, we retrieve the link to the Wikipedia page and retrieve the Q-item through an API call to the WikiMedia API. DBPedia supports less languages than WikiData and YAGO, and the information of an entity is not always present in all languages. To ensure that the maximum performance by DBPedia is achieved, a fallback mechanism is implemented, where if an entity is not encountered in the local DBPedia version, an attempt to retrieve the English version is made. This significantly improved the scores of the model. Ideally, we would want to input entities into the system and bypass the entity recognition system, as we know the inputs are entities. Although DBPedia has this functionality, it is only available for English and works very poorly when applied to other languages. Therefore, the entity recognition component is used but a simple string matching filter is used to ensure no completely inaccurate guesses are made by the system due to language coverage issues.

3.2.2. WikiData

WikiData is a knowledge base created by the WikiMedia foundation, containing roughly 97 million entities in more than 300 languages.[3] For querying WikiData we use the SPARQL endpoint for the WikiData API, using the 'EntitySearch' feature and retrieve the Q-items for the returned entities. We only retrieve the first entity from a list of responses, and set the language for each of the queries, depending on the language of the entity.

3.2.3. YAGO

YAGO (Suchanek et al., 2007) is another knowledge base that builds on Wikidata, with the latest version ,YAGO4, containing roughly 64 millions entities at the time of writing. YAGO stores facts in RDF format and uses logical constraints to increase the coherence of the knowledge base, for example by making sure entities can not be persons and places at the same time. [4] For querying YAGO, a similar approach to the one used for WikiData is used, using the SPARQL endpoint of YAGO for querying, providing the language of entities depending on the language the entities are in.

3.3. Comparison

As the ParlaMint corpus is a very large corpus that consists of multiple languages and alphabets, annotating a large set of entities for entity linking is not very feasible. In order to obtain a proxy for the performance of the models on Parlamint, and evaluate their performance on different languages, a baseline test was performed on the names of local politicians from ten countries, extracted from WikiData using membership querying. (Query can be found in Appendix 1). This method of obtaining gold standard for the entity linking process was chosen over manual annotation of ParlaMint entities, as it provides us with high quality Named Entity names that do not contain the noise discussed previously, such as aliasing. However, as the Named Entities used are all members of parliament in their respective countries, we feel that these entities provide an accurate representation of (part of) the ParlaMint corpus and therefore the results obtained for the samples of local politicians should provide a good proxy on the results of the entity linkers on the real ParlaMint data, albeit an ideal case.

For the comparison experiment, we collected 100 members of parliament from ten countries together with their Q-item through a membership query performed on WikiData. We then ran all three systems on the 100 members from parliament, and reported their accuracy for the countries respectively. For the politicians, only people that started in office from 01-01-2014 onward were selected, to be in line with the time period of the ParlaMint project.

This test was conducted to obtain scores of the systems in 'ideal' conditions, with correctly written full names and with minimal ambiguity. This allows us to later manually 'distort' these entities to investigate the effect of aliasing while maintaining gold standard links. It also provides us with a means of comparing the performance of the entity linking systems across different languages, allowing us to analyse whether the performance differs between different languages or language families.

[3] https://www.wikidata.org/wiki/Wikidata:Statistics

[4] https://yago-knowledge.org/getting-started

3.4. Lemmatization

In order to study the effect of lemmatization on the performance of the three systems, we measure the amount of inflection for all entity types by comparing how many times the original string is equal to the lemma, to get in indication of the amount of inflection for different countries. To gain a more detailed understanding, we selected twenty frequent entities such as Angela Merkel and Donald Trump from the ParlaMint corpus and selected inflections by finding entities that contain these entities as substring. Thus for each entity we obtain a list of variants of that name. For each of these variants, we run WikiData, as this was the best performing model in the comparison, and calculate the overall precision by weighting the scores of each variant by the amount of times they occur, to get a more realistic indication of the effect of lemmatization when applied to individual entities.

3.5. Aliasing

Because entities are often unambiguous within a local context, aliasing can occur, following Grice's Maxim of quantity. That is, given a situation in which an entity is known to the participants in for example a debate, referencing this person by surname provides the appropriate amount of information to successfully disambiguate that person in that context, without the superfluous addition of the first name when this is not required. However, when attempting to link individual terms, this phenomenon becomes problematic, as it increases the ambiguity of an entity.

To study the effect of aliasing on the performance of the three models, we set up an experiment where we only use surnames for the entity linking process. We use the local politicians collected for the 'ideal' scenario here, as these can be easily changed and we can readily generate the gold standard for them. We decided to limit the experiment to five countries, namely The Netherlands, Belgium, France, Poland and the United Kingdom. For each of these countries, we select 10 entities and remove their first names. For example 'Margaret Thatcher' becomes 'Thatcher'. We then evaluate the performance of the three models on these lists of surnames and report the scores.

3.5.1. Temporal De-Aliasing Algorithm

The method used in this paper is similar to the method used in Gottipati and Jiang (2011). We start with an entity E and a list of discovered variants V. At the start, this record only contains E itself, $V = \{E\}$. Now we find all other Named Entities in the document with the same type as E, and if they contain E as a substring, they are added to V. To maximise the number of discovered variants, we also introduce a temporal parameter in the algorithm, which determines how many debates 'around' the mention of the entity we consider for discovering variants. After this procedure, we obtain the variant v^* from V that occurred the most in the considered documents (excluding E itself). This entity v^* is then used as the query to the knowledge base.

3.5.2. Restricting Considered Entities

Apart from the temporal based approach, we also experiment with the usage of the metadata available for the ParlaMint corpora. In this version we make use of the lists of members of parliament available for a specific country. For a named entity found in the text, we compare it to the database of parliamentary members of that country using a simple cosine similarity score between character n-grams of the surnames of the target entity and the knowledge base. We use character two and three grams for encoding the entities into vectors. As some entities might not be present in the metadata of that particular country (such as ministers from different countries) we also consider ministers from other countries if no compatible match is found within the metadata of the country itself. If no entity has a high enough similarity threshold, we report it as a NIL entity. Because the performance of this method partly relies on the entities selected for the linking (i.e. only selecting local entities will prevent the step of using metadata from different countries to have an effect), we take a balanced sample of local politicians and entities referenced in multiple countries (the 'international entities'), instead of using the names from local politicians from the WikiData membership query. For both categories, we select ten entities at random.

3.6. Code

Our code is available at `https://github.com/RubenvanHeusden/LRECMultilingualEntityLinkingCode`

4. Results

In this section the results to the experiments posed in Section 3 are presented in the order that they are discussed above.

4.1. Comparison

Table 1 shows the results of running DBPedia, WikiData and YAGO on the automatically retrieved local politicians. One thing that can be noticed immediately is the high performance of the WikiData system on the task. One obvious reason for this is the fact that the entities were extracted from the WikiData knowledge base, and therefore the system is more likely to get the entities correct. However, some mistakes are still made by the WikiData system. Further inspection of the results showed that this was almost entirely due to ambiguous names, which caused WikiData to link with incorrect entities, for example 'James Morris' being linked to a researcher instead of a politician for the United Kingdom, or 'Sophie Hermans' being linked to a researcher instead of the correct politician for the Netherlands.

For DBPedia, the scores are on par with WikiData for a few countries such as NL and FR, but fall behind for

Country	DBPedia	WikiData	YAGO
NL	0.97	**0.98**	0.56
DE	0.58	**0.94**	0.60
FR	0.95	**0.97**	0.95
CZ	0.31*	**0.95**	0.87
HU	0.75	**0.90**	0.73
EN	0.74	**0.87**	0.78
IT	0.18*	0.95	**0.97**
IS	0.67*	**1.00**	0.85
DK	0.69	**0.96**	0.79
TR	0.52	**0.97**	0.71
Mean	0.74	**0.94**	0.73

Table 1: Accuracy of DBPedia, WikiData and YAGO on 100 local politicians from 8 countries. (∗ signifies that the model either did not support the language, or the language was not properly recognized. These countries were also not considered for the mean of the system performance).

most other countries. There are several reasons for this lower performance, the main reason being the inability of the system to recognize entities. If an entity is not recognized or only partially recognized, a correct link cannot be made. To eliminate the effect of mistakes in the recognition of DBPedia, we have also used the 'search' API. However, this API is only available in English, and although it can sometimes link entities in other languages, this is by no means guaranteed. Furthermore, although Czech and Italian are reportedly supported by DBPedia, the API was not able to retrieve resources in those languages. For YAGO the main problem is also that the system does not recognize the entity present, and thus returning a NIL result.

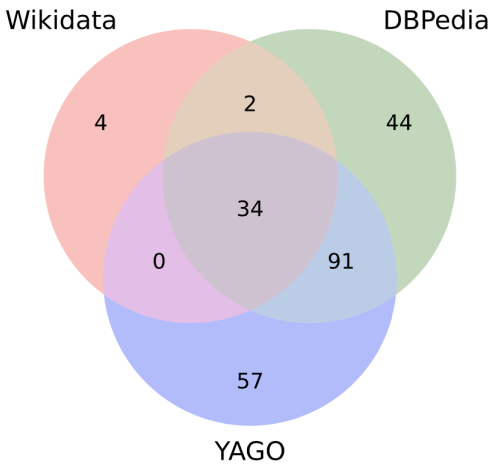

Figure 1: Venn diagram showing the overlap between the mistakes of the 3 systems (excluding IT, IS and CZ).

Figure 1 shows the distribution of errors between the three systems. The first observation that can be made is that there are only a few instances in which WikiData makes a mistakes that the other two systems did not make. However, we do see that in the cases of both DBPedia and YAGO, the system makes mistakes that the others two systems do not make, more often. DBPedia and YAGO also overlap on a large number of cases, showing that these systems are quite similar not only in scores (as seen in Table 1) but also in the type of mistakes they make. The overlap between all three systems shows that when WikiData makes a mistake, the other two systems almost always also make that mistake. In the majority of cases where all three systems made the same mistake, this concerned the miss classification of an entity, rather than the system outputting a NIL prediction.

If we compare the performances of the systems across different languages, we can see that WikiData is quite stable across different languages, with English being the worst performing language. This can be partially explained by the fact that English is the most prevalent language on WikiPedia, and thus more cases of ambiguity arise than for other languages, a hypothesis supported by the types of mistakes made by WikiData. For DBPedia and YAGO there is a bit more variance across languages, with YAGO scoring relatively low on NL and DE, as well as on TR. DBPedia scores higher on NL, but also scores relatively low on DE and TR, suggesting a gap in the coverage of entities in those languages for the two systems. However, these results are on ideal cases in which the name is in canonical form, and the full name is used. It does give as an indication of the relative performances of the systems on the languages in ParlaMint. Next we will investigate what happens when these ideal conditions are not met, in the cases of the presence of inflections name variants.

4.2. Lemmatization

In this section, the results of lemmatization are presented, with several examples being given, and a detailed analysis of lemmatization being made for the PER entities of seven countries. In Table 2 several examples of the names of people being inflected are shown. Inspection of the lemmas found that among the countries that inflect words most often are Polish, Czech and Latvian. With for example Dutch and English having virtually no inflections, something that is in line with the intuition about the morphologies of these languages.

Entity	**Inflections**
Angela Merkel	Angeli Merkel Merkelova
Donald Tusk	Donaldem Tuskiem Donaldzie Tusku Donaldowi Tuskowi

Table 2: Examples of inflections of popular entities in different languages in the Polish language.

As can be seen from Table 3, the amount of lemmatization various greatly from country to country, as well as from type to type. Especially the MISC entity type is often changed after lemmatization. This is not unexpected, as the MISC entity type can contain a great variety of entities, and thus these might be lemmatized more often.

	LOC	MISC	ORG	PER
LV	-	-	0.87	0.60
TR	0.52	-	0.72	0.45
IS	0.67	0.80	0.64	0.41
CZ	0.77	0.38	0.65	0.41
PL	0.87	-	0.76	0.36
HR	0.62	0.91	0.69	0.36
SI	0.76	0.91	0.75	0.34
IT	0.06	-	0.13	0.26
FR	0.40	0.18	0.35	0.24
BE	0.09	0.42	0.26	0.15
HU	-	0.26	0.18	0.15
LT	0.24	0.50	0.88	0.05
DK	0.12	0.61	0.41	0.04
NL	0.04	0.41	0.10	0.03
ES	0.01	0.08	0.04	0.02
BG	0.09	0.73	0.68	0.01
GB	0.00	0.06	0.01	0.00

Table 3: Fraction of the unique entities in each subcorpus of ParlaMint that changed after lemmatization. NaN values indicate the category was not present in that subcorpus. Sorted on the PER entity type.

Surprisingly, organisations also get lemmatized frequently. Examples of this include 'Partij voor de Dieren' being lemmatized to 'Partij voor de Dier' in The Netherlands, and 'east midlands trains' being lemmatized to 'east midlands train' in the United Kingdom, removing the plural 's'. This suggests using the lemmatized version of organisations might actually be harmful to the performance entity linking models on those entities. Investigating the PER entity type it can be seen that countries such as Latvia, Turkey and Icelandic have entities that are lemmatized often, and thus we expected these countries to benefit most from using lemmas for entity linking.

In Table 4, the results of lemmatization are shown on the names of twenty international PER entities for seven countries when linked using WikiData. It can be seen that for PL, CZ, HR and IS, the lemmatization has a clear positive effect on the scores of the EL system, showing that for these languages lemmatization is beneficial. For NL and BG however, the usage of lemmatization has a negative effect on performance, especially for BG. This is most likely due to the fact that these languages do not inflect words often, and thus lemmatization might 'correct' entities that do not need to be corrected. An example of this for NL would be the lemmatization of 'Edith Schippers' into 'Edith Schipper', where a correct entity is lemmatized into an incorrect one.

	Percentage of entities recognized	
Country	Before lemmatization	After lemmatization
PL	0.33	0.53
CZ	0.37	0.67
HR	0.29	0.74
IS	0.67	0.75
LV	0.16	0.24
BG	0.77	0.40
NL	0.91	0.89

Table 4: Accuracy of the WikiData system on a set of 20 entities, before and after lemmatization.

To conclude this research question, the usage of lemmatization has a significant positive impact on several languages with a large number of inflections, such as PL, CZ and HR. For languages with a low number of inflections, such as BG and NL, the lemmatization has no effect, and for BG, the performance is actually severely hampered by the unnecessary use of lemmatization.

4.3. Aliasing

When evaluating the systems on the manually aliased names, it was found that all three systems failed to recognize persons only mentioned by their surname, achieving a score of zero for all tested countries. However, it is important to mention that in the case of DBPedia, the system does not return any entity, while in the case of WikiData and YAGO, the systems often returned a 'family name' entity for the surname or a reference to a disambiguation page.

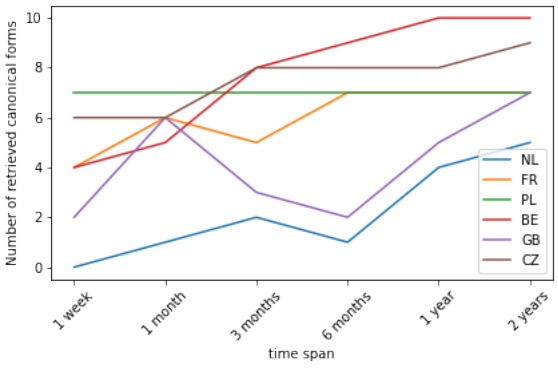

Figure 2: Results of applying the aliasing algorithm with various time spans for six countries (the time span is from both sides, so '1 month' means 1 month earlier and 1 month later).

In Table 2 the results of applying the time-based dealiasing algorithm with various settings for the temporal granularity are shown. The y-axis represents the

total number of entities that was correctly resolved, out of a maximum of ten entities. For most countries, the amount of correctly de-aliased mentions increases as the time span parameter is increased with the exception of Poland, for which the number of resolved entities remains the same. The drops in the number of resolved entities can be explained by the fact that over a certain time period, for ambiguous entities, another incorrect variant might be more popular than the correct variant, causes a drop that is later resolved as the time span is increased. This will largely depend on the chosen entity, as less ambiguous names will not have this problem to the same extent.

4.3.1. Constricted Entity Disambiguation

Below are the results of applying the disambiguation method that only considers entity present in the speaker metadata of the specific country, or the speaker metadata of the other countries. As can be seen from Table 5, the performance of a simple EL system using string similarity performs relatively poor when considering only local entities. This is not surprising, as the samples are a mix of local and international figures. However, for some countries the scores for using only local politicians are also low for the local politicians group. This is the case in Belgium, where it was found that most entities from the sample were in fact not parliamentary actors. In the case of using speaker metadata from multiple parliaments, the performance of the simple model on all countries is increased, suggesting this approach definitely has some merit over the approach only using local entities.

Country	Only local	Multiple Parliaments
NL	0.55	0.80
FR	0.45	0.60
PL	0.35	0.60
BE	0.20	0.35
GB	0.40	0.55
CZ	0.35	0.45

Table 5: Results of applying the de-aliasing approach based on ParlaMint speaker metadata, with using only metadata from the country itself, and metadata on members of parliament from other countries. For each country, 20 entities were evaluated.

To conclude this research question, we found that the simple time based de-aliasing method we used is already quite effective for some cases in the de-aliasing of names, although the limitations of the method are also clear. This does provide us with some insights into the problem of aliasing, and possibilities for future work on more complicated methods. One interesting possibility could be to extend the idea of the constricted entity linking method, and incorporate the usage of the linked metadata present in some of the corpora, with links to Wikipedia, Twitter or other external sources. These sources can then be used to provide more context surrounding the entity, to provide a model with more information in the case of ambiguous entities, a method often used within the field of Entity Linking.

5. Discussion & Future Work

In future work, the approaches used for alleviating the effects of aliasing could be refined, by for example using context from debates for the surnames and using methods such as BERT other Transformer based models to score entities. For the analysis of the lemmatization effects, the lemmatizers that each country employed themselves were used. Without detailed knowledge of the language and the software used, there is no way of assessing the quality of these lemmatizers. This might cause differing results for the lemmatization of certain countries. Although this work only deals with the PER entities present in the ParlaMint corpus, it can also be extended to the other entity types present in the corpus. The problems of lemmatization and aliasing also exist for these entity types, albeit in slightly different forms and severities. For organization names, aliasing will most likely take the form of abbreviations of names, which could be resolved through the usage of local context, possibly combined with a list of abbreviations for large organisations. In the case of locations, the main challenge in linking the entities (apart from lemmatization) is the ambiguity arising from different locations having the same name. This could possibly be resolved by only considering locations within the country of the parliamentary debate, or giving higher weights to locations within that country.

6. Conclusion

In this paper we investigated the performance of three entity linking systems on data from the ParlaMint corpus, and we found that the WikiData system performed the best overall for the local politicians, although all systems performed relatively well. Through investigation of the ParlaMint dataset, we found that for certain languages, entities are often inflected or entities are referred to by aliases. These phenomena create noise in the dataset, and are problematic for creating entity links for all entities in ParlaMint. We investigated the effect of lemmatization on the entities in the dataset by using the automatically generated lemmas of the entities and comparing the performance of WikiData on entities before and after lemmatization. We found that for PL, CZ and HR, lemmatization had a big effect, while in particular for BG and NL the effects were negligible or it actually hampered performance, in the case of BG. Thus for some languages, lemmatization can have a profound positive effect on the performance of entity linking systems, although one must be careful in choosing which languages to use it for, as to not harm the performance of the model by lemmatizing unnecessarily. Finally, we investigated the effect of aliasing on the ability of models to properly link entities, by manually aliasing ten ground truth politicians for

five languages. We found that it severely inhibited the models from finding the correct entities. Through the usage of a simple heuristic using corpus statistics and term occurrence in files, a significant portion of names could be resolved, although the simplicity of the heuristic also introduces errors concerning ambiguity, leaving an interesting opportunity for future work.

7. Language Resource References

Erjavec, Tomaž and Ogrodniczuk, Maciej and Osenova, Petya and Ljubešić, Nikola and Simov, Kiril and Grigorova, Vladislava and Rudolf, Michał and Pančur, Andrej and Kopp, Matyáš and Barkarson, Starkaur and Steingrímsson, Steinþór and van der Pol, Henk and Depoorter, Griet and de Does, Jesse and Jongejan, Bart and Haltrup Hansen, Dorte and Navarretta, Costanza and Calzada Pérez, María and de Macedo, Luciana D. and van Heusden, Ruben and Marx, Maarten and Çöltekin, Çağrı and Coole, Matthew and Agnoloni, Tommaso and Frontini, Francesca and Montemagni, Simonetta and Quochi, Valeria and Venturi, Giulia and Ruisi, Manuela and Marchetti, Carlo and Battistoni, Roberto and Sebők, Miklós and Ring, Orsolya and Darġis, Roberts and Utka, Andrius and Petkevičius, Mindaugas and Briedienė, Monika and Krilavičius, Tomas and Morkevičius, Vaidas and Bartolini, Roberto and Cimino, Andrea and Diwersy, Sascha and Luxardo, Giancarlo and Rayson, Paul. (2021). *Linguistically annotated multilingual comparable corpora of parliamentary debates ParlaMint.ana 2.1*.

8. Bibliographical References

Botha, J. A., Shan, Z., and Gillick, D. (2020). Entity linking in 100 languages. *arXiv preprint arXiv:2011.02690*.

De Cao, N., Wu, L., Popat, K., Artetxe, M., Goyal, N., Plekhanov, M., Zettlemoyer, L., Cancedda, N., Riedel, S., and Petroni, F. (2021). Multilingual autoregressive entity linking. *arXiv preprint arXiv:2103.12528*.

Färber, M., Ell, B., Menne, C., and Rettinger, A. (2015). A comparative survey of dbpedia, freebase, opencyc, wikidata, and yago. *Semantic Web Journal*, 1(1):1–5.

Gottipati, S. and Jiang, J. (2011). Linking entities to a knowledge base with query expansion. Association for Computational Linguistics.

McNamee, P., Mayfield, J., Lawrie, D., Oard, D. W., and Doermann, D. (2011). Cross-language entity linking. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 255–263.

Mendes, P. N., Jakob, M., García-Silva, A., and Bizer, C. (2011). Dbpedia spotlight: shedding light on the web of documents. In *Proceedings of the 7th international conference on semantic systems*, pages 1–8.

Pappu, A., Blanco, R., Mehdad, Y., Stent, A., and Thadani, K. (2017). Lightweight multilingual entity extraction and linking. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, pages 365–374.

Pillai, S. G., Soon, L.-K., and Haw, S.-C. (2019). Comparing dbpedia, wikidata, and yago for web information retrieval. In *Intelligent and Interactive Computing*, pages 525–535. Springer.

Sil, A., Kundu, G., Florian, R., and Hamza, W. (2018). Neural cross-lingual entity linking. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Suchanek, F. M., Kasneci, G., and Weikum, G. (2007). Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web*, pages 697–706.

A. Appendix

SPARQL query to retrieve local politicians

```
SELECT ?item ?itemLabel ?group ?groupLabel
?district ?districtLabel ?term ?termLabel ?start ?end
WHERE
{
  ?item p:P39 ?statement .
  ?statement ps:P39/wdt:P279* wd:%s ; pq:P580 ?start .
  OPTIONAL { ?statement pq:P2937 ?term }
  OPTIONAL { ?statement pq:P582  ?end }
  OPTIONAL { ?statement pq:P768  ?district }
  OPTIONAL { ?statement pq:P4100 ?group }
  FILTER((!BOUND(?end) || ?end > NOW())
  && (?start >  "2014-01-01T00:00:00+00:00"^^xsd:dateTime) )
  SERVICE wikibase:label { bd:serviceParam wikibase:language "[AUTO_LANGUAGE],en". }
}
ORDER BY ?start ?end
```

Listing 1: Example of Named Entity XML tag

Trump
Macron
Salvini
Putin
Kennedy
Berlusconi
Merkel
Juncker
Cameron
Obama
Blair
Thatcher
Stalin
Barnier
Hitler
Johnson
Tusk
Churchill
Timmermans
Hollande

Table 6: Entities used for the lemmatization Research Question