Modelling Voting Behaviour within the European Parliament

Ross Chadwick [1]

[1](Department of Computer Science, Vrije Universiteit Amsterdam)

June 2018 (Draft)

1 Introduction

Beginning in 2001, a European regulation was passed providing public access rights to all documents produced by the European Parliament, Council and Commission [1]. The information published by these European institutions provides rich resources for many scientific disciplines. Ranging from insights into large-scale politics [2, 3, 4], human-behaviour analysis [5], social-network analysis [6, 7, 8, 9] and natural language processing [10, 11].

Although such information is publicly available, it is mostly provided in an unstructured manner and in a variety of formats. Increased data interoperability can be achieved if European Union data transitions to 5 star Linked Open Data (LOD) [12]. This transition would open new opportunities for large-scale, complex analysis into the European Union which are not currently feasible without proprietary data-sets, or building custom data aggregation systems.

Within this research, we will focus on the topic of voting within the European Parliament. Specifically, by modelling behaviour at the lowest level of abstraction; individual Members of Parliament's (MEPs) votes. Modelling vote relations on an individual basis will allow for in-depth analysis into how MEPs form stances on vote topics. This could in-turn, provide a base for higher abstractions, such as political group voting habits.

We therefore propose to model the legislative process of a dossier passing through the European Parliament, in conjunction with all related activities, documents, votes and other relevant metadata as LOD. The knowledge graph produced will then be linked to existing knowledge graphs to further enrich the LOD cloud, within the domain of the European Parliament.

Furthermore, machine learning methods will be explored in an attempt to predict future voting behaviour of MEPs, based on past votes and available metadata within the graph. This stage will provide a proof-of-concept link prediction application, illustrating a potential use-case for such a knowledge graph. The techniques used will then be compared to give an indication of potential future directions needed to improve model accuracy in such a domain.

2 Research Questions

1. **Can modelling the European Parliament legislative system as LOD provide an enhanced platform for analytics, compared to existing data-sets?**
   Modelling a machine-readable knowledge graph with references to many external LOD resources should provide new insights into parliamentary data in ways that was not feasible with existing data formats. This question will be assessed by providing in-depth statistical analysis of the types of relations and data available, compared to existing data-sets.

2. **Can MEP's future votes be predicted based on their past voting record and metadata?**
   Votes will be modelled as relations in the produced Linked Data graph, allowing the usage of existing link prediction techniques to predict voting behaviour of individual MEPs. Multiple predictive models will be built and compared as a proof-of-concept, demonstrating the effectiveness of each approach in this context.
   If this question can be answered positively, an analysis into which metadata best serve as predictors could be formed.

3 Background

3.1 The Current State of Open European Union Data

There exists many initiatives focused on utilising and improving European Union data. Very few however touch on the subject of voting outcomes and even fewer do this in a free and open manner. VoteWatch [13] and Parltrack [14] are two notable initiatives aiming to improve the accessibility of European Parliament voting data. While VoteWatch is highly comprehensive, data-sets are only available under a paid subscription model. On the contrary, Parltrack is fully free and open source. The Parltrack data-sets however lack coherent structuring in their current form; requiring proficient data aggregation skills to be utilised effectively. Both initiatives therefore fall short in different ways when it comes to catering towards the wider academic community.

LOD related to the European Union is still largely unavailable. There exists one official initiative by the European Council [15] which does in-fact provide voting data from the council, as well as metadata of requested council documents. There is however no accompanying initiative on behalf of the parliament, which is what this research project will aim to produce. LinkedEP, an independent LOD initiative published by Aggelen, Hollink et al. [16] provides structured data on European Parliament plenary debates. In the case of LinkedEP, it has been shown that such information has a substantial demand, with 7500 unique queries executed within the first seven months of the data publication [17].

Many relevant resources related to the European Parliament do already have a

presence on the Semantic Web. For example, DBPedia contains many triples on European Parliament political groups and MEPs. Resource properties do however vary greatly in quality and quantity and should not be relied on for accurate, wholesome information. MEP resources do occasionally contain useful properties such as relations within national political groups as well as European Parliament political groups [18].

3.2 Machine Learning over Knowledge Graphs and Link Prediction Techniques

Modern machine learning techniques have shifted towards favouring data with as little manual feature engineering as possible. More specifically, deep learning has allowed data scientists to utilise a layered model approach to derive relevant features from their data in an end-to-end fashion. Wilcke, Bloem et al. have argued that it is often problematic to acquire data in a raw, unaltered form and that the distributed knowledge graphs of the LOD cloud could provide the ideal data model for such tasks [19].

RDF2Vec, proposed by Ristoski, Paulheim [20] is one such approach for utilising LOD graphs within traditional machine learning models. The authors adapt language modelling methods for unsupervised feature extraction to RDF based graphs, using Weisfeiler-Lehman Subtree RDF Graph Kernels and graph walks to extract graph sub-structures. In short, entities and relations in a graph are converted into a sequence of entities, which can then be treated like sentences in training neural language models. The resulting vector space embedding indicates the semantic similarity between entities in a graph. Their experimental results demonstrate that this approach outperforms traditional feature generation approaches for simple tasks such as classification and regression. Testing this approach in more complex applications, such as link prediction is still actively being explored [].

Although RDF2Vec has been shown to be an effective approach to learning over large scale graphs, it does not operate in a fully end-to-end fashion. Important information may be lost in the transformation step from RDF to feature vectors, which is not recoverable for fine-tuning feature extraction.

One approach which is truly an end-to-end model is the Relational Graph Convolutional Network (RGCN), as proposed by Schlichtkrull, Kipf, Bloem et al. [21]. The RGCN is an adaption of the traditional GCN, allowing usage in the domain of knowledge graphs. With this approach every layer of the model, including feature extraction, can be fine-tuned based on the next layer's error signal. Compared to RDF2Vec, this can greatly reduce the chance of sub-optimal feature extraction.

4 Methodology

4.1 European Parliament Linked Open Data

1. All relevant information on procedures in Parliament will be scraped and converted into linked data, with an accompanying ontology to assist in the inference of relations. Parltrack provides in-depth data-sets [22] under the Open Database License, scraped from the official Parliament website, reducing the need to build custom scrapers.

2. Votes will be modelled as relations between MEPs and the legislative document voted on. This graph structure will form the basis of the data-set for link prediction, as will be further explained in section 4.2.

3. The ontology will be modelled with immutability in mind, to mitigate the likelihood of triples requiring updates.

4. Resources will be linked to their external counterparts where available (DBPedia, LinkedEP, GeoNames, European Council, JRC-Names):

    (a) DBPedia - MEPs, political groups and locations will be linked to their existing DBPedia Resource (DBR).

    (b) LinkedEP - MEPs will be linked to their resource. LinkedEP AgendaItem resources will be linked to their corresponding procedural activity.

    (c) GeoNames - All locational resources, will be linked to their GeoNames resource for a wider range of geographical information on dossiers and people.

    (d) European Council - When the European Council has had voting involvement within a procedure, the relevant resources will be linked to.

    (e) JRC-Names [23] - MEP resources will be linked to their JRC-Names resource where available, to provide access to variations of their names. This should greatly improve the ability of querying over MEPs in a multilingual fashion.

5. Textual resources will be annotated with mentioned topics using tools such as DBPedia Spotlight [24], GATE [25] or possibly a custom topic modelling solution.

6. Lastly, a system will be built to allow for regular automatic extension of the knowledge graph, so that it remains relevant to current politics.

4.2 Prediction of Future Votes

1. The knowledge graph as described in section 4.1 will be converted into a specialised data-set, specifically catered towards machine learning applications. This will then be utilised to train and test the proposed proof-of-concept link prediction models. The data-set will be released publicly

under an open license, in the hope that researchers can use it to benchmark future models in a standardised manner.

2. Multiple proof-of-concept link prediction-based multi-label classification models will be built, to be trained and tested on the produced data-set. The models will output newly predicted vote relations (for, against, abstain) between MEPs and legislative documents, with a metric of certainty indicating the likelihood of each vote relation type.
   Likely approaches to be explore are:

   (a) RDF2Vec
   (b) Relational Graph Convolutional Networks (RGCN)

3. Hyperparameters will be tuned for each model, and the results from each optimised setup will be compared, using metrics such as ROC AUC and overall accuracy.

4. A comparative analysis will be made, to investigate variations in the tested approaches.

5. If the results seem to answer the research question to an acceptable level, it will be possible to make a short analysis into what types of metadata appear most influential in shaping an MEP's vote.

6. The section will be concluded and insights into future directions, such as incorporating multi-modal (Specifically natural language) data will be mentioned.

5 Potential Problems

The second research question relies on the assumption that an MEP's vote is influenced by their past voting patterns and other relevant metadata. However researchers such as Hix et al. [4] discuss the fact that an MEP's vote may not always reflect their actual opinion on legislation, but is in-fact influenced by their affiliated political group and cross-group politics. If the proposed predictive models are not capable of discovering patterns on a political group level as well as on an individual level, we may observe very low accuracies.

6 Preliminary Work

Work is already well underway modelling the ontology and producing triples of some aspects that will be considered. A prototype implementation of the data mining tool which produces said triples and updates a SPARQL endpoint with them is fully functional and outputs around 20 million triples (including basic OWL2-RL inferences). The tool currently relies on the Parltrack data-sets and does not do any scraping itself. It will be explored if this is a reasonable way

to continue, or if the Parltrack scraping tools will be forked and customised to allow complete control over the scraping process.

Figure 1 demonstrates an example visualisation of some of the triples currently available for an MEP resource. In this example, the target resource is in fact a DBPedia resource for Terry Reintke, which has inherited properties through an $owl:sameAs$ relation to the internal MEP resource. The figures following show the current properties that the most notable resources in the graph contain.

Firstly we can see that this MEP has some basic properties associated (2a), such as birth date, gender and name. The MEP then has a membership (2b), showing that they represent Germany within the Verts/ALE political group as of 01-07-2014. The absence of an end date indicates that this membership is still active. Next we see a reaction (3a) that this MEP has made towards a voting round, in which they voted "For" as indicated by the reaction having type $epv:VoteFor$. We can then see some basic information on the voting round in question (3b). Looking further we can follow the relations of the vote, all the way to the dossier (procedure) that it applies to (4a, 4b, 4c).
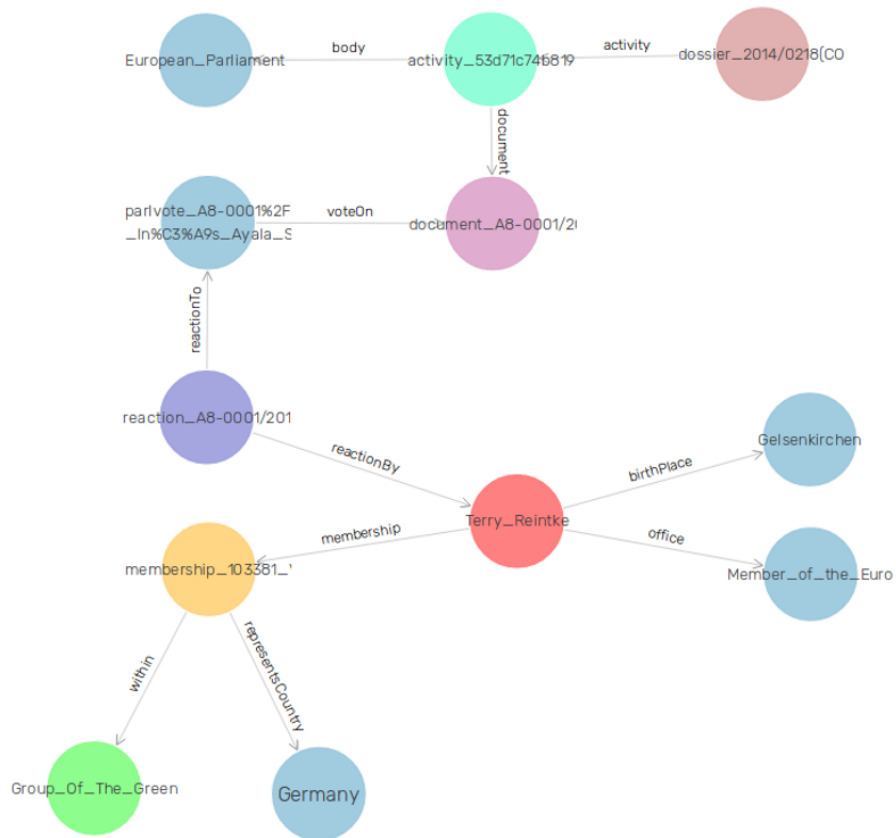
Figure 1: Example MEP (Terry Reintke) and a subset of their relations

(a) MEP Properties

(b) MEP Membership Properties

reaction_A8-0001/2015_103381_For

◇ reaction_A8-0001/2015_103381_For

Types:
`epv:VoteFor`
`owl:Thing`

(a) MEP's vote "For"

parlvote_A8-0001%2F2015_-_ln%C3%A9s_Ayala_Sender_-_R%C3%A9solution_L%C3%A9gislative

◇ parlvote_A8-0001%2F2015_-_ln%C3%A9s_Ayala_Sender_-_R%C3%A9solution_L%C3%A9gislative

Types:
`owl:Thing`

epv:voteTitle
A8-0001/2015 - Inés Ayala Sender - Résolution législative

xsd:anyURI
http://www.europarl.europa.eu/RegData/seance_pleniere/proces_verbal/2015/02-11/votes_nominaux/xml/P8_PV(2015)02-11(RCV)_XC.xml

(b) The vote in parliament which the MEP reacted to

document_A8-0001/2015

◇ document_A8-0001/2015

Types:
`epv:Document`
`epv:Report` `owl:Thing`
epv:documentTititle
A8-0001/2015

rdf:type
committee report tabled for plenary, 1st reading/single reading

xsd:anyURI
http://www.europarl.europa.eu/sides/getDoc.do?type=REPORT&mode=XML&reference=A8-2015-0001&language=EN

(a) The report which was the basis of the vote

activity_53d71c74b819f204 00000_committee_report_ ed_for_plenary,_1st_readi ngle_reading

◇ activity_53d71c74b819f2042b000000_comn

Types:
`epv:Activity`
`owl:Thing`
xsd:date
2015-01-07

(b) The parliamentary activity of which the document belongs to

dossier_2014/0218(COD)

◇ dossier_2014/0218(COD)

Types:
`epv:Dossier`
`owl:Thing`
epv:dossierTitle
Cross-border exchange of information on road safety related traffic offences

xsd:anyURI
http://www.europarl.europa.eu/oeil/popups/ficheprocedure.do?reference=2014/0218(COD)&l=en

(c) The dossier which the activity was in regards to

References

[1] Regulation (ec) no 1049/2001 of the european parliament and of the council. `http://www.europarl.europa.eu/RegData/PDF/r1049_en.pdf`.

[2] Bjorn Hoyland, Jean-François Godbout, Emanuele Lapponi, and Erik Velldal. Predicting party affiliations from european parliament debates. In *LTCSS@ACL*, 2014.

[3] Derek Greene and James P. Cross. Unveiling the political agenda of the european parliament plenary: A topical analysis. `https://arxiv.org/pdf/1505.07302.pdf`, 2015.

[4] Simon Hix, A. Noury, and Gérard Roland. Voting patterns and alliance formation in the european parliament. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 364 1518:821–31, 2009.

[5] Stephen A. Meserve, Daniel Pemstein, and William T. Bernhard. Political ambition and legislative behavior in the european parliament. `https://pdfs.semanticscholar.org/eea5/1373d2c2165f925d102d529f190ad23f3709.pdf`, 2007.

[6] Darko Cherepnalkoski and Igor Mozetic. Retweet networks of the european parliament: evaluation of the community structure. *Applied Network Science*, 1:2, 2016.

[7] Darko Cherepnalkoski, Andreas Karpf, Igor Mozetic, and Miha Grcar. Cohesion and coalition formation in the european parliament: Roll-call votes and twitter activities. In *PloS one*, 2016.

[8] Darko Cherepnalkoski and Igor Mozetic. A retweet network analysis of the european parliament. *2015 11th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*, pages 350–357, 2015.

[9] Attila Kovács. Political networks in the european parliament: Network analysis of the 2013 common agricultural policy reform. `https://kgk.uni-obuda.hu/sites/default/files/10-Attila-Kovacs.pdf`.

[10] Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. `http://homepages.inf.ed.ac.uk/pkoehn/publications/europarl-mtsummit05.pdf`, 2005.

[11] Najeh Hajlaoui, David Kolovratník, Jaakko Väyrynen, Ralf Steinberger, and Dániel Varga. Dcep -digital corpus of the european parliament, 2014.

[12] 5 star open data. `http://5stardata.info`.

[13] Votewatch eu. `http://www.votewatch.eu/`.

[14] Parltrack. `http://parltrack.euwiki.org/`, .

[15] Council of the european union - open data initiatives. `http://www.consilium.europa.eu/en/general-secretariat/corporate-policies/transparency/open-data/`.

[16] Linkedpolitics - plenary debates of the european parliament as linked open data. `http://linkedpolitics.ops.few.vu.nl`.

[17] Astrid van Aggelen, Laura Hollink, Max Kemman, Martijn Kleppe, and Henri Beunders. The debates of the european parliament as linked open data. *Semantic Web*, 8:271–281, 2017.

[18] Example dbpedia mep resource. `http://dbpedia.org/resource/Peter_van_Dalen`.

[19] Xander Wilcke, Peter Bloem, and Victor De Boer. The knowledge graph as the default data model for learning on heterogeneous knowledge. *Data Science*, 1(1-2):39–57, 12 2017. ISSN 2451-8484. doi: 10.3233/DS-170007.

[20] Petar Ristoski and Heiko Paulheim. Rdf2vec: Rdf graph embeddings for data mining. In *International Semantic Web Conference*, 2016.

[21] Michael Sejr Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks. In *ESWC*, 2018.

[22] Parltrack database schema. `http://parltrack.euwiki.org/dumps/schema.html`, .

[23] Maud Ehrmann, Guillaume Jacquet, and Ralf Steinberger. Jrc-names: Multilingual entity name variants and titles as linked data. *Semantic Web*, 8:283–295, 2017.

[24] Joachim Daiber, Max Jakob, Chris Hokamp, and Pablo N. Mendes. Improving efficiency and accuracy in multilingual entity extraction. In *Proceedings of the 9th International Conference on Semantic Systems (I-Semantics)*, 2013.

[25] Hamish Cunningham, Diana Maynard, Kalina Bontcheva, and Valentin Tablan. Gate : an architecture for development of robust hlt applications. 2002.