An Ontology Design Pattern for Describing Personal Data in Privacy Policies

Harshvardhan J. Pandit, Declan O'Sullivan, and Dave Lewis

ADAPT Centre, Trinity College Dublin, Dublin, Ireland
{harshvardhan.pandit|declan.osullivan|dave.lewis}@adaptcentre.ie

Abstract. Privacy laws such as the General Data Protection Regulation (GDPR) specify several obligations involving personal data. A privacy policy is a document that provides information for legal compliance on how personal data is collected, used, stored, and shared, which is essential for understanding their privacy implications. Approaches such as the UsablePrivacy project that extract information from the text of the privacy policy need to structure it in a manner suitable for machine processing. Semantic web has been proven to be suitable to represent this knowledge as a set of queryable concepts and relationships. However, there is a large overlap between different projects and approaches targeting the privacy policy that does not take advantage of the significant similarity of its underlying information. We present an ontology design pattern to aid these efforts in representing and modelling information related to personal data within a privacy policy. The pattern aims to assist the existing ecosystem of machine-based approaches for interpretation and visualisation of privacy policies by providing a common structured representation to ease modelling and sharing of related information.

Keywords: Ontology Design Pattern, Personal Data, Privacy Policy, GDPR

1 Motivation & Scope

A privacy policy provides disclosure of information regarding the collection, usage, storage, and sharing of personal data as governed by territorial laws [11]. The privacy policy is most commonly presented as a monolithic text document. It has been repeatedly proven to be difficult to read and understand[1] for the users [2,3,5].

To remedy this, there have been several approaches towards making interpretation of privacy policies easier for end users. 'Terms of Service; Didn't Read'[2] is a community driven approach to summarise privacy policies in the interest of user awareness. Some recent approaches use machine-learning to interpret the contents of the privacy policy and to present their analysis to the user in a visual

[1] The impact of GDPR on the readability of privacy policies is yet to be determined
[2] https://tosdr.org/index.html

format. This allows the approaches to scale with the ever-increasing and changing nature of services and their privacy policies. Relevant approaches regarding this are the UsablePrivacy project[3] [6] and the PrivacyGuide project [9,10].

The privacy policy provides information associated with personal data such as its collection, usage, sharing, and storage along with other information such as the provision of various rights required by law. The General Data Protection Regulation (GDPR) [1], which is a European law for data protection, requires specification of this information in privacy policies for compliance. This provides commonality with respect to the mandatory information specified by laws which is provided in privacy policies. Machine-processing approaches require this information to be structured in a machine-readable format that can aid in the automation of processes. Approaches that target the same set of information have to deal with the same set of data - in this case the information regarding personal data provided in privacy policies. This commonality of underlying information (within privacy policies) can be represented using a common vocabulary for its expression. This will aid the different approaches through sharing of extracted information from privacy policies, while also making it possible to compare the efficiency of different methods in extracting this information.

Using a semantic web based approach provides a way to define such knowledge in the form of concepts and relationships with the freedom for them to be expanded and connected based on requirements. The UsablePrivacy project already uses such an approach involving semantic web ontologies to represent its underlying information about the categorisation of sentences within a privacy policy [6]. In addition to such machine-based approaches, other work involving the information presented within a privacy policy will also benefit from such structuring of information based on semantic web technologies.

With this as our motivation, we present an ontology design pattern (ODP) for modelling the information related to personal data within a privacy policy. This ODP provides a way to express the personal data and the information associated with it as a set of concepts and relationships which can be incorporated into a larger semantic web ontology.

In terms of scope, we limit to expressing information related to personal data provided explicitly within a privacy policy. Other relevant information within a privacy policy but not part of the pattern is discussed as future work at the end of this paper. In terms of relevant work, we are not aware of any similar approaches for modelling of information within a privacy policy towards creating an ontology design pattern.

The ODP is based on an investigation of privacy policies from Airbnb Ireland[4] and Twitter [5], with archived copies made available[6] in case of changes to the policy in future. While we also evaluated other privacy policies for their structure and content, we specify these two as being our primary use-cases for the purpose

[3] https://www.usableprivacy.org
[4] https://www.airbnb.ie/terms/privacy_policy
[5] https://twitter.com/en/privacy
[6] https://opengogs.adaptcentre.ie/harsh/privacy-policy-dashboard/

of this work. An investigation of information within the privacy policy provided by Airbnb Ireland is available online [7] but is not part of this paper's contribution.

The rest of this paper is structured as follows: Section 2 provides a description of the pattern with an example provided in Section 3. Section 4 concludes the paper with a discussion regarding future work.

2 Pattern Description

2.1 Competency Questions

The pattern aims to answer the following competency questions:

1. What personal data is collected? e.g. email
2. Does the data have a category? e.g. contact information
3. What was its source? e.g. user
4. How is it collected? e.g. given by user, automated
5. What is it used for? e.g. creating an account, authentication and verification
6. How long is it retained for? e.g. 90days after account deletion
7. Who is it shared with? e.g. name of partner organisation(s)
8. What is the legal basis? e.g. given consent, legitimate use
9. What processes/purposes was the data shared for? e.g. analytics, marketing
10. What is the legal type of third party? e.g. processor, controller, authority

The pattern does not consider questions related to the provision of GDPR rights. While these questions are relevant, they are directly related to the data subject (or user), and are common to all instances of personal data. They are better represented in the model of the privacy policy rather than as an instance of personal data. We provide them here for brevity, with a further discussion on this provided in the future work section:

11. How can personal data be rectified or corrected?
12. How can personal data be deleted or removed?
13. How can a copy of the personal data be obtained?
14. How can personal data be transferred to another party?
15. How can information about the personal data be obtained?

The pattern uses the GDPRtEXT[7] and GDPRov[8] ontologies for defining concepts relevant to the GDPR. GDPRov is an ontology for describing the provenance of consent and personal data lifecycles using GDPR relevant terminology, and is an extension of PROV-O and P-Plan. GDPRtEXT provides definitions of concepts and terms used within the text of the GDPR using SKOS.

The pattern is available online along with its documentation[8] and has been submitted to the ontology design patterns collaborative wiki[9].

[7] `http://openscience.adaptcentre.ie/privacy-policy/personalise/demo/policy.html`

[8] `https://openscience.adaptcentre.ie/projects/privacy-policy/design-pattern/`

[9] `http://ontologydesignpatterns.org/wiki/Submissions:PrivacyPolicyPersonalData`

2.2 Concepts & Relationships

A visualisation of the pattern is presented in Fig. 1, and was created using the yEd graph editor[10] with the Graffoo [4] palette.

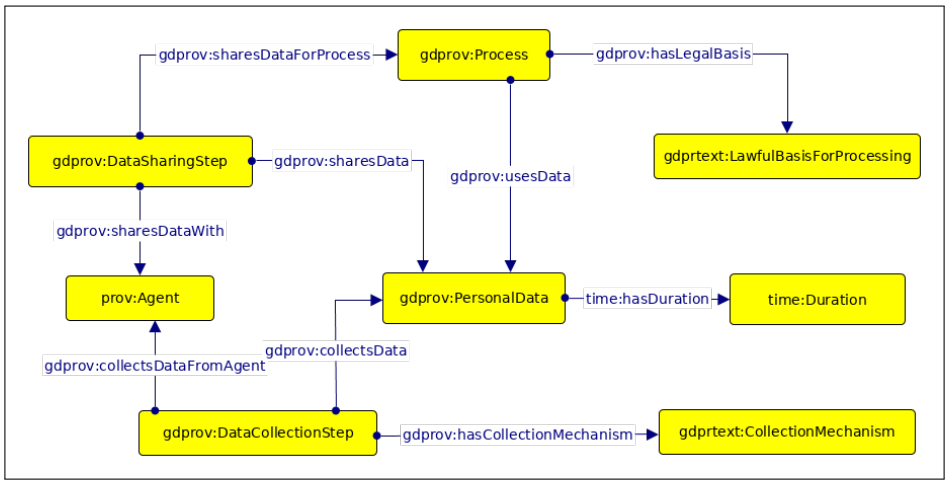

Fig. 1. Personal data pattern illustrated using Graffoo [4]

Personal Data *PersonalData* represents an instance of personal data, such as an *email address*, which is described in the privacy policy. It is defined as an instance of *gdprov:PersonalData*. Privacy policies often group related instances of personal data in broader categories such as *contact information* for representing *email* and *phone number*. To represent such a grouping in the pattern, the category can be represented as a subclass of *PersonalData* using *rdfs:subClassOf*, with its instances representing individual personal data items. In this case, *contact information* would be a subclass of *gdprov:PersonalData* with *email* and *phone number* being its instances.

$$PersonalDataCategory \sqsubseteq PersonalData \tag{1}$$

Data Collection Data is collected through a *gdprov:DataCollectionStep*, and is represented using the property *gdprov:collectsData*. The data provider is represented using *prov:Agent* through the property *gdprov:collectsDataFromAgent*.

Apart from the source, the privacy policy may also mention the particular collection mechanism used for data collection. This is represented using the *gdprov:hasCollectionMechanism* property, where the collection mechanism

[10] `https://www.yworks.com/products/yed`

is represented by a suitable subclass of *gdprtext:CollectionMechanism*, such as *gdprtext:GivenByUser* or *gdprtext:AutomatedCollection*.

$$DataCollectionStep \sqsubseteq \geq 1 collectsData.PersonalData \tag{2}$$

$$DataCollectionStep \sqsubseteq \geq 1 collectsDataFromAgent.Agent \tag{3}$$

$$PersonalData \sqsubseteq \forall hasCollectionMechanism.CollectionMechanism \tag{4}$$

Data Retention The retention of personal data informs how long it would be stored for. This is represented using the Time Ontology in OWL[11], which is the W3C recommendation for describing temporal concepts. The retention period is represented using the property *time:hasDuration* with the range being an instance of *time:Duration*. This information is arbitrary and may be missing or in an un-representable format within the privacy policy such as "retained for as long as necessary" where the information is difficult to represent.

$$PersonalData \sqsubseteq \forall hasDuration.Duration \tag{5}$$

Data Usage & Processing Data Usage, also termed as Processing, is the use of personal data for some purpose as specified within the privacy policy. This can vary in terms of granularity from a comparatively simple step such as sending an email to a more abstract process such as marketing which encompasses several steps and processes. The pattern therefore uses *gdprov:Process*, which is a subclass of *p-plan:Plan*, to define a process which can contain one or more steps and processes. The property *gdprov:usesData* is used to represent the use of personal data within a process.

$$Process \sqsubseteq \geq 1 usesData.PersonalData \tag{6}$$

Legal Basis for Data Usage Every use of personal data within a process must have a legal basis under the GDPR. Examples of such legal basis defined within GDPRtEXT include consent, legitimate interest, compliance with the law, and performance of contract. To represent this, the pattern uses the property *gdprov:hasLegalBasis* with the range *gdprtext:LawfulBasisForProcessing*. Since every data use must have at least one legal basis, this provides the axiom:

$$Process \sqsubseteq \geq 1 hasLegalBasis.LawfulBasisForProcessing \tag{7}$$

Data Sharing The sharing of data involves the entity the data is shared with, the purposes for sharing, and their legal basis. This is represented within the pattern through the use of *gdprov:DataSharingStep* and the property *gdprov:sharesData*. The entity the data is shared with is represented using the *gdprov:sharesDataWith* property with the domain as *gdprov:DataSharingStep* and the range as a type of

[11] `https://www.w3.org/TR/owl-time/`

gdprov:Agent, such as another Data Controller, Data Processor, or an Authority. The purpose of sharing is represented using *gdprov:Process* and the property *gdprov:sharesDataForProcess* to model the data being used in that process after sharing. The legal basis of processes for which the data is shared is represented using *gdprov:hasLegalBasis* as specified earlier. Since it is mandatory to inform who the data is being shared with, along with its intended purposes, and the specific legal obligation, we have the following axioms:

$$DataSharingStep \sqsubseteq \geq 1 sharesData.PersonalData \tag{8}$$

$$DataSharingStep \sqsubseteq \geq 1 sharesDataWith.Agent \tag{9}$$

$$DataSharingStep \sqsubseteq \geq 1 sharesDataFor.Process \tag{10}$$

3 Example Use-Case

We present here an example use-case of the pattern for depicting personal data from Airbnb Ireland's privacy policy. The use-case was chosen for its generality in terms of being common to other privacy policies as well as ease of understanding for users.

The use-case concerns the 'email address' specified as personal data within the privacy policy, which is provided by the user. It is used to "provide, improve, and develop platform services", which is specified as a process with the legal basis of legitimate interest. It is shared with the 'Payments Controller' entity for 'Identity Verification' process which has a legal basis of 'contract fulfilment'.

The example use-case is illustrated in Fig. 2 using Graffoo [4] and shows the classes, properties, and instances. The corresponding code is presented in Listing. 1 using the Turtle[12] notation for RDF. The answers to the competency questions corresponding to the use-case are provided below.

1. What personal data is collected: Email Address
2. Does the data have a category: Account Information
3. What was its source: User
4. How is it collected: Given by user
5. What is it used for: Platform Services, Payments
6. How long is it retained for: indefinitely (no end duration)
7. Who is it shared with: Payments Controller
8. What is the legal basis: Legitimate Interest, Contract
9. What processes/purposes was the data shared for: Identity Verification
10. What is the legal type of third party: Data Controller

[12] `https://www.w3.org/TR/turtle/`

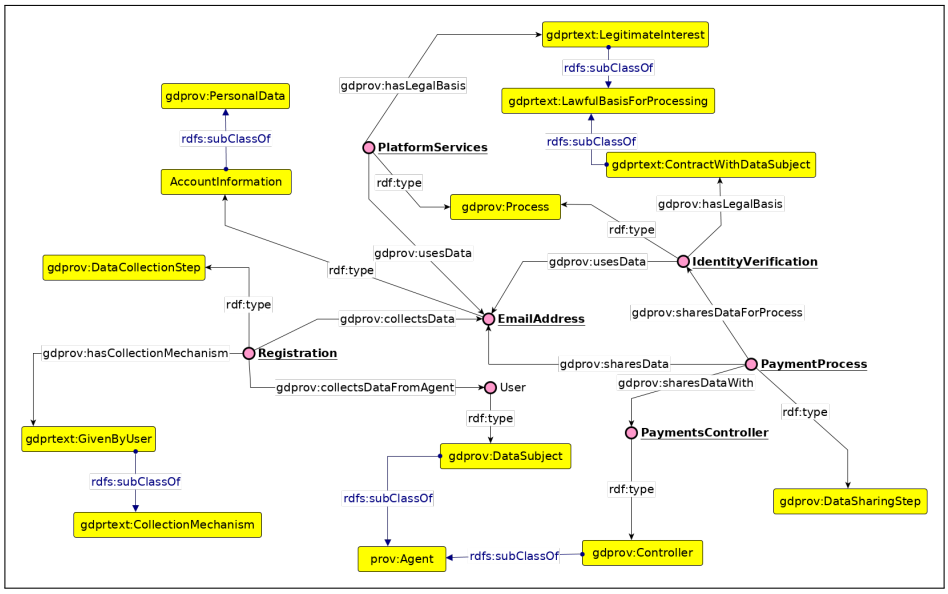

Fig. 2. Example Use-case illustrated using Graffoo [4] and showing collection, storage, usage, and sharing of Email Address as Personal Data

4 Conclusion

This paper presents an ontology design pattern for representing information associated with personal data in the context of a privacy policy. More specifically, it allows modelling and representation of collection, usage, storage, and sharing of personal data along with the associated processes and entities, as well as their legal basis. Concepts and relationships within the pattern are defined using the previously published GDPRov [8] and GDPRtEXT [7] ontologies. For defining the duration of storage, the pattern uses the Time ontology in OWL.

The paper provides an use-case of the pattern based on an real-world privacy policy to reflect its suitability. The use-case captures one instance of personal data and the information related to it within the privacy policy. The depicted usage of ontology design patterns provides motivation for adoption in approaches related to the use of information within a privacy policy. This allows sharing of information through a common representation for related activities such as summarising, visualisation, analytics, or determining compliance using information contained within privacy policies.

Based on the intended motivation, the pattern provides a way to share the relevant information regarding personal data, and provides further avenues for research regarding similar patterns or meta-patterns related to privacy policies.

Future Work

We consider our work an initial effort towards consolidating information within privacy policies. Using the pattern to reflect information from several distinct real-world privacy policies will demonstrate its feasibility and applicability in real-world scenarios. This presents a challenge as the pattern currently assumes the presence of all required information which may not be the case for some use-cases, particularly where interpretations of information are ambiguous. However, capturing such ambiguities through a meta-pattern can possibly aid in flagging them for review by legal experts.

In addition to the above, the pattern faces other challenges for the modelling of information it aims to represent. For example, it is not clear what level of abstraction should be represented in the pattern regarding concepts such as storage and sharing. Should there be a *DataStorageStep* which can be further annotated to represent various pieces of information relating to the storage of personal data? Abstractions can help to represent different storage duration and formats for the same instance of personal data, such as storing the actual data for 6 months while a (pseudo-)anonymised copy is stored for 2 years. However, tacking on such abstractions in to the pattern can make it rigid (in terms of modelling) and complex. More work needs to be undertaken to evaluate whether such abstractions are necessary in the pattern, and how they should be represented.

Another challenge is the representation of storage duration (or retention period). Concrete values such as 6 months or 2 years can be represented using appropriate ontologies, but ambiguous statements are difficult to represent using such ontologies. An example of this is the statement "data may be stored for as long as necessary..." in which there is no end to the duration for storage. Representing this as a *time:Duration* instance is problematic as there is no clear method to represent its end period. Not defining an end period is also not a solution due to the open world assumption. Our approach towards solving this issue is to abstract the storage activity as described earlier. However, we are open for other approaches and solutions towards this problem.

The privacy policy contains more information than is reflected by the pattern. To represent this additional set(s) of information, larger (combinations of) patterns and ontologies will be needed to model and represent all the relevant information and context. This is especially relevant for GDPR as it mandates the inclusion of information regarding its various rights, which is presented through the privacy policy.

Some of this information was presented in this paper as additional competency questions. These help evaluate information regarding how the personal data can be changed (rectified), deleted, and obtained (download a copy). Additionally, GDPR allows the data subject to change their consent, thereby affecting the processes involving personal data. Capturing this information is essential towards quantifying the privacy policies into machine-readable data, with the paper demonstrating the suitability of ODP for this task.

```
@prefix dct: <http://purl.org/dc/terms/> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix xml: <http://www.w3.org/XML/1998/namespace> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix gdprov:
    <http://purl.org/adaptcentre/openscience/ontologies/gdprov#> .
@prefix gdprtext:
    <http://purl.org/adaptcentre/openscience/ontologies/GDPRtEXT#> .
@prefix : <http://example.com/personaldata#> .

:PaymentProcess a gdprov:DataSharingStep ;
    rdfs:label "Payment Process"^^xsd:string ;
    gdprov:sharesData :EmailAddress ;
    gdprov:sharesDataForProcess :IdentityVerification ;
    gdprov:sharesDataWith :PaymentsController .

:PlatformServices a gdprov:Process ;
    rdfs:label "Provide, Improve, and Develop Platform"^^xsd:string ;
    gdprov:hasLegalBasis gdprtext:LegitimateInterest ;
    gdprov:usesData :EmailAddress .

:Registration a gdprov:DataCollectionStep ;
    rdfs:label "Registration for new users"^^xsd:string ;
    gdprov:collectsData :EmailAddress ;
    gdprov:collectsDataFromAgent :User ;
    gdprov:hasCollectionMechanism gdprtext:GivenByUser .

:AccountInformation a rdfs:Class, owl:Class ;
    rdfs:label "Account Information of an User"^^xsd:string ;
    rdfs:subClassOf gdprov:PersonalData .

:IdentityVerification a gdprov:Process ;
    rdfs:label "Identity Verification"^^xsd:string ;
    gdprov:hasLegalBasis gdprtext:Contract ;
    gdprov:usesData :EmailAddress .

:PaymentsController a gdprov:Controller,
        prov:Agent ;
    rdfs:label "Payments Controller"^^xsd:string .

:User a gdprov:DataSubject,
        prov:Agent ;
    rdfs:label "User of Service"^^xsd:string .

:EmailAddress a :AccountInformation,
        :PersonalData ;
    rdfs:label "Email Address"^^xsd:string .
```

Listing 1: Example Use-case in Turtle format presenting Email Address as an instance of personal data along its collection, storage, and sharing

Acknowledgements

This work is supported by the ADAPT Centre for Digital Content Technology which is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund.

References

1. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). Official Journal of the European Union L119, 1–88 (May 2016), http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L:2016:119:TOC
2. Contissa, G., Docter, K., Lagioia, F., Lippi, M., Micklitz, H.W., Pa\lka, P., Sartor, G., Torroni, P.: Claudette Meets GDPR: Automating the Evaluation of Privacy Policies Using Artificial Intelligence (2018)
3. Fabian, B., Ermakova, T., Lentz, T.: Large-scale Readability Analysis of Privacy Policies. In: Proceedings of the International Conference on Web Intelligence. pp. 18–25. WI '17, ACM, New York, NY, USA (2017), http://doi.acm.org/10.1145/3106426.3106427
4. Falco, R., Gangemi, A., Peroni, S., Shotton, D., Vitali, F.: Modelling owl ontologies with graffoo. In: Presutti, V., Blomqvist, E., Troncy, R., Sack, H., Papadakis, I., Tordai, A. (eds.) The Semantic Web: ESWC 2014 Satellite Events. pp. 320–325. Springer International Publishing, Cham (2014)
5. Jensen, C., Potts, C.: Privacy Policies As Decision-making Tools: An Evaluation of Online Privacy Notices. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. pp. 471–478. CHI '04, ACM, New York, NY, USA (2004), http://doi.acm.org/10.1145/985692.985752
6. Oltramari, A., Piraviperumal, D., Schaub, F., Wilson, S., Cherivirala, S., Norton, T., Russell, N., Story, P., Reidenberg, J., Sadeh, N.: PrivOnto: A semantic framework for the analysis of privacy policies. Semantic Web 9(2), 185–203 (Jan 2018), http://www.medra.org/servlet/aliasResolver?alias=iospress&doi=10.3233/SW-170283
7. Pandit, H.J., Fatema, K., O'Sullivan, D., Lewis, D.: GDPRtEXT - GDPR as a Linked Data Resource. p. 14. Heraklion, Crete, Greece (2018)
8. Pandit, H.J., Lewis, D.: Modelling Provenance for GDPR Compliance using Linked Open Data Vocabularies. In: Proceedings of the 5th Workshop on Society, Privacy and the Semantic Web - Policy and Technology (PrivOn2017) (PrivOn) (2017), http://ceur-ws.org/Vol-1951/#paper-06
9. Tesfay, W.B., Hofmann, P., Nakamura, T., Kiyomoto, S., Serna, J.: I Read but Don'T Agree: Privacy Policy Benchmarking Using Machine Learning and the EU GDPR. In: Companion Proceedings of the The Web Conference 2018. pp. 163–166. WWW '18, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland (2018), https://doi.org/10.1145/3184558.3186969
10. Tesfay, W.B., Hofmann, P., Nakamura, T., Kiyomoto, S., Serna, J.: PrivacyGuide: Towards an Implementation of the EU GDPR on Internet Privacy Policy Evaluation. In: Proceedings of the Fourth ACM International Workshop on Security

and Privacy Analytics. pp. 15–21. IWSPA '18, ACM, New York, NY, USA (2018), http://doi.acm.org/10.1145/3180445.3180447
11. Wilson, S., Schaub, F., Dara, A.A., Liu, F., Cherivirala, S., Giovanni Leon, P., Schaarup Andersen, M., Zimmeck, S., Sathyendra, K.M., Russell, N.C., B. Norton, T., Hovy, E., Reidenberg, J., Sadeh, N.: The Creation and Analysis of a Website Privacy Policy Corpus. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1330–1340. Association for Computational Linguistics, Berlin, Germany (Aug 2016), http://www.aclweb.org/anthology/P16-1126