# Overview of the First Content Selection Challenge from Open Semantic Web Data

**Nadjet Bouayad-Agha**[1]
**Gerard Casamayor**[1]
**Leo Wanner**[1,2]
[1]DTIC, University Pompeu Fabra
[2]Institució Catalana de Recerca i Estudis Avançats
Barcelona, Spain
firstname.lastname@upf.edu

**Chris Mellish**
Computing Science
University of Aberdeen
Aberdeen AB24 3UE, UK
c.mellish@abdn.ac.uk

## Abstract

In this overview paper we present the outcome of the first content selection challenge from open semantic web data, focusing mainly on the preparatory stages for defining the task and annotating the data. The task to perform was described in the challenge's call as follows: given a set of RDF triples containing facts about a celebrity, select those triples that are reflected in the target text (i.e., a short biography about that celebrity). From the initial nine expressions of interest, finally two participants submitted their systems for evaluation.

## 1 Introduction

In (Bouayad-Agha et al., 2012), we presented the NLG challenge of content selection from semantic web data. The task to perform was described as follows: given a set of RDF triples containing facts about a celebrity, select those triples that are reflected in the target text (i.e., a short biography about that celebrity). The task first required a data preparation stage that involved the following two subtasks: 1) *data gathering and preparation*, that is, deciding which data and texts to use, then downloading and pairing them, and 2) *working dataset selection and annotation*, that is, defining the criteria/guidelines for determining when a triple is marked as selected in the target text, and producing a corpus of triples annotated for selection.

There were initially nine interested participants (including the two organizing parties). Five of which participated in the (voluntary) triple annotation rounds.[1] In the end, only two participants submitted their systems:

**UA:** Roman Kutlak, Chris Mellish and Kees van Deemter. Department of Computing Science , University of Aberdeen, Scotland (UK).

**UIC:** Hareen Venigalla and Barbara Di iEugenio. Department of Computer Science, University of Illinois at Chicago (USA).

Before the presentation of the baseline evaluation of the submitted systems and the discussion of the results (Section 4), we outline the two data preparation subtasks (Sections 2 and 3). In Section 5, we then sketch some conclusions with regard to the achievements and future of the content selection task challenge. More details about the data, annotation and resources described in this overview, as well as links for downloading the data and other materials (e.g., evaluation results, code, etc.) are available on the challenge's website.[2]

## 2 Data gathering and preparation

We chose Freebase as our triple datastore.[3,4] We obtained the triple set for each person in the Turtle format (ttl) by grepping the official Freebase RDF dump released on the 30th of December 2012 for all triples whose subject is the person's URI; certain meta-data and irrelevant triples (i.e., triples with specific namespaces such as "base" or "common") have been filtered out.

Each triple set is paired with the person's summary biography typically available in Wikipedia, which consists of the first paragraph(s) preceding the page's table of contents[5]

Our final corpus consists of 60000+ pairs, all of which follow two restrictions that are supposed to

[1]We would like to thank Angelos Georgaras and Stasinos Konstantopoulos from NCSR (Greece) for their participation in the annotation rounds.

[2]http://www.taln.upf.edu/cschallenge2013/

[3]http://www.freebase.com

[4]For a comparison between Freebase and DBPedia, see http://wiki.freebase.com/wiki/DBPedia.

[5]For example, the first four paragraphs in the following page constitute the summary biography of that person: http://en.wikipedia.org/wiki/George_Clooney.



*Proceedings of the 14th European Workshop on Natural Language Generation*, pages 98–102,
Sofia, Bulgaria, August 8-9 2013. 

maximize the chances of having interesting pairs with sufficient original and selected input triples for the challenge. Firstly, the number of unique predicates in the input ttl must be greater than 10. The number 10 is estimated based on the fact that a person's nationality, date and place of birth, profession, type and gender are almost always available and selected, such that we need a somewhat large set to select content from in order to make the task minimally challenging. Secondly, the Wikipedia-extracted summary biography must contain more than 5 anchors and at least 20% of the available anchors, where an anchor is a URI in the text (i.e., external href attribute value in the html) pointing to another Wikipedia article which is directly related to that person. Given that most Freebase topics have a corresponding DBPedia entity with a Wikipedia article, anchors found in the introductory text are an indicator of potential relevant facts available in Freebase and are communicated in the text. In other words, the anchor threshold restriction is useful to discard pairs with very few triples to annotate. We found this criterion more reliable than the absolute length of the text which is not necessarily proportional with the number of triples available for that person.

## 3 Working Dataset selection and annotation

The manual annotation task consisted in emulating the content selection task of a Natural Language Generation system, by marking in the triple dataset associated with a person the triples predicated in the summary biography of that person according to a set of guidelines. We performed two rounds of annotations. In the first round, participants were asked to select content for the same three celebrities. The objectives of this annotation, in which five individuals belonging to four distinct institutions participated, were 1) for participants to get acquainted with the content selection task envisaged, the domain and guidelines, 2) to validate the guidelines, and 3) to formally evaluate the complexity of the task by calculating inter-annotator agreement. For the latter we used free-marginal multi-rater Kappa, as it seemed suited for the annotation task (i.e. independent ratings, discrete categories, multiple raters, annotators are not restricted in how they distribute categories across cases) (Justus, 2005). We obtained an average Kappa of 0.92 across the three pairs for the 5 annotators and 2 categories (selected, not selected), which indicates a high level of agreement and therefore validates our annotation guidelines.

Our objective for the second round of annotations was to obtain a dataset for participants to work with. In the end, we gathered 344 pairs from 5 individuals of 5 distinct institutions. It should be noted that although both rounds of annotations follow the anchor restriction presented in Section 2, the idea to set a minimum number of predicates for the larger corpus of 60000+ pairs came forth after analysing the results of the second round and noting the data sparsity in some pairs. In what follows, we detail how the triples were presented to human annotators and what were the annotation criteria set forth in the guidelines.

### 3.1 Data presentation

A machine-readable triple consists of a subject which is a Freebase machine id (mid), a predicate and an object which can either be a Freebase mid or a literal, as shown in the following two triples:

```
ns:m.0dvld
  ns:people.person.spouse_s
    ns:m.02kknf3 .

ns:m.0dvld
  ns:people.person.date_of_birth
    "1975-10-05"^^xsd:datetime .
```

Triples were transformed into a human-readable form. In particular, each mid in object position (e.g., 02kknf3) was automatically mapped onto an abbreviated description of the Freebase topic it refers to. Thus, the triples above have been mapped onto a tabular form consisting of (1) predicate, (2) object description, (3) object id, and (4) object types (for literals):

```
(1) /people/person/spouse_s
(2) "1998-11-22 - Jim Threapleton -
     2001-12-13 - Marriage -
     Freebase Data Team - Marriage"
(3) /m/02kknf3

(1) /people/person/date_of_birth
(2) value
(3) "1975-10-05"
(4) "datetime"
```

For each triple thus presented, annotators were asked to mark 1) whether it was selected, 2) in which sentence(s) of the text did it appear, and 3) which triples, if any, are its coreferents. Two triples are coreferent if their overlap in meaning is such that either of them can be selected to represent the content communicated by the same text

fragment and as such should not count as two separate triples in the evaluation. Thus, the same text might say *He is probably best known for his stint with heavy metal band Godsmack* and *He has also toured and recorded with a number of other bands including Detroit based metal band Halloween 'The Heavy Metal Horror Show'* ..., thus referring in two different sentences to near-equivalent triples `/music/artist/genre ``Heavy metal"` and `/music/artist/genre ``Hard rock"`.

### 3.2 Annotation criteria

Annotators were asked to first read the text carefully, trying to identify propositional units (i.e., potential triples) and then to associate each identified propositional unit with zero, one or more (coreferent) triples according to the following rules:

*Rule 1.* One cannot annotate facts that are not predicated and cannot be inferred from predicates in the text. In other words, all facts must be grounded in the text. For example, in the sentence *He starred in Annie Hall*, the following is predicated: `W.H.has_profession actor` and `W.H. acted_in film Annie Hall`. The former fact can be inferred from the latter. However, the following is not predicated: (1) `Person has_name W.H.`, (2) `W.H. is Male`, and (3) `W.H. is Person`.

*Rule 2.* In general, one can annotate more generic facts if they can be inferred from more specific propositions in the text, but one cannot annotate specific facts just because a more general proposition is found in the text. In the example *He was a navigator*, we can mark the triples `Person has_profession Sailor` as well as `Person has_profession Navigator` (we would also mark them as coreferent). However, given the sentence *He was a sailor*, we cannot mark the triple `Person has_profession Navigator`, unless we can infer it from the text or world knowledge.

*Rule 3.* One can annotate specific facts from a text where the predicate is too vague or general if the facts can be inferred from the textual context, from the available data, or using world knowledge. This rule subsumes four sub-cases:

*Rule 3.1.* The predicate in the proposition is too vague or general and can be associated with multiple, more specific triples. In this case, do not select any triple. In the example *Film A was a great commercial success*, we have several triples associating the celebrity with Film A, as director, actor, writer, producer and composer and none of them with a predicate resembling "commercial success". In this case there are no triples that can be associated with the text.

*Rule 3.2.* The predicate in the proposition is too vague or general, but according to the data there is just one specific triple it can be associated with. In this case, select that triple. In the example *Paris released Confessions of an Heiress*, the term `released` could be associated with *authored*, *wrote* or *published*. However, there is only one triple associating that subject with that object, which matches one of the interpretations (i.e., *authoring*) of the predicate. Therefore that triple can be selected.

*Rule 3.3.* The predicate in the proposition is too vague or general, but one or more specific triples can be inferred using world knowledge. In this case, select all. The sentence *He is also a jazz clarinetist who performs regularly at small venues in Manhattan*, can be associated with the available triples `W.H. profession Clarinetist` and `W.H. music/group_member/instruments_played Clarinet`, even though for this latter triple the person being in a group is not mentioned explicitly. However, this can be inferred from basic world knowledge.

*Rule 3.4.* The predicate in the proposition is too vague or general, but one or more specific triples can be inferred using the textual context. In this case, select all. In the example *By the mid-1960s Allen was **w**riting and directing films ... Allen often **s**tars in his own films ... Some of the best-known of his over 40 films are Annie Hall (1977)* ..., the relations of the person with the film `Annie Hall` are that of writer, director and actor, as supported by the previous text. Therefore we would annotate facts stating that the person wrote, directed and starred in `Annie Hall`. However, we wouldn't annotate composer or producer triples if they existed.

*Rule 4.* A proposition can be associated with multiple facts with identical or overlapping meanings. In the example, *Woody Allen is a musician*, we have the triples `W.H occupation musician` and `W.H profession musician`, which have near

identical meanings. Therefore, we mark both triples and indicate that they co-refer. The sentence *Woody Allen won prize as best director for film Manhattan*, on the other hand, can be associated with non-coreferring triples `W.H won prize` and `W.H. directed Manhattan`.

*Rule 5.* If the text makes reference to a set of facts but it does not enumerate them explicitly, and there is no reason to believe it makes reference to any of them in particular, then do not annotate individual facts. Thus, sentence *Clint Eastwood has seven children* does not warrant marking each of the seven children triples as selected, given that they are not enumerated explicitly.

*Rule 6.* If the text makes a clear and unambiguous reference to a fact, do not annotate any other facts, even though they can be inferred from it. In other words, as explained in Rule 1, all annotated triples must be grounded in the text. In the sentence *For his work in the films Unforgiven (1992) and Million Dollar Baby (2004), Eastwood won Academy Awards for Best Director and Producer of the Best Picture*, we can infer from world knowledge that the celebrity was nominated prior to winning the award in those categories. However, the text makes a clear reference only to the fact that he won the award and there is no reason to believe that it is also predicating the fact that the celebrity was nominated.

## 4 Baseline evaluation

Briefly speaking, the UA system uses a general heuristic based on the cognitive notion of *communal common ground* regarding each celebrity, which is approximated by scoring each lexicalized triple (or property) associated with a celebrity according to the number of hits of the Google search API. Only the top-ranked triples are selected (Kutlak et al, 2013). The UIC system uses a small set of rules for the conditional inclusion of predicates that was derived offline from the statistical analysis of the co-occurrence between predicates that are about the same topic or that share some shared arguments; only the best performing rules tested against a subset of the development set are included (Venigalla and Di Eugenio, 2013).

For the baseline evaluation, we used the development set obtained in the second round annotation (see Section 3). However, we only consider pairs obtained during the second round annotation that 1) follow both restrictions presented in Section 2, and 2) have no coreferring triples. This last restriction was added to minimize errors because we observed that annotators were not always consistent in their annotation of triple coreference.[6] We therefore considered 188 annotations from the 344 annotations of the development set. Of these, we used 40 randomly selected annotations for evaluating the systems and 144 for estimating a baseline that only considers the top 5 predicates (i.e., the predicates most often selected) and the type-predicate.[7].

| | Baseline | UIC | UA |
|---|---|---|---|
| Precision | 49 | 64 | 47 |
| Recall | 67 | 50 | 39 |
| F1 | 51 | 51 | 42 |

Table 1: Baseline evaluation results (%)

The evaluation results of the three systems (baseline, UIC and UA) are presented in Table 1. The figures in the table were obtained by comparing the triples selected and rejected by each system against the manual annotation. The performance of the baseline is quite high. The UA system based on a general heuristic scores lower than the baseline, whilst the UIC system has a better precision than the baseline, albeit a lower recall. This might be due, as the UA authors observe in their summary (Venigalla and Di Eugenio, 2013), to "the large number of predicates that are present only in a few files ...[which] makes it harder to decide whether we have to include these predicates or not."

## 5 Conclusions

We have given an overview of the first content selection challenge from open semantic web data, focusing on the rather extensive and challenging technological and methodological work involved in defining the task and preparing the data. Unfortunately, despite agile participation in these early

[6]Type-predicate triples were filtered out of the annotated files in the development set whilst they were included in the large corpus made available to the candidates. Therefore, we added type-predicate triples in the development set a posteriori for this evaluation. These type-predicate triples might be coreferring with other triples, say `ns:m.08rd51 ns:type.object.type ns:film.actor` and `ns:m.08rd5 people/person/profession "Actor" /m/02hrh1q .` Nonetheless, this was not taken into account in the evaluation.

[7]The top 5 predicates were (in descending order of frequency): music track, film actor, profession, date of birth and nationality

preparatory stages, the number of submitted systems was limited. Both of the presented systems were data-intensive in that they usedeither a pool of textual knowledge or the corpus of triple data provided by the challenge in order to select the most relevant data.

Unlike several previous challenges that involve more traditional NLG tasks (e.g., surface realization, referring expression generation), content selection from large input semantic data is a relatively new research endeavour in the NLG community that coincides with the rising interest in statistical approaches to NLG and dates back, to the best of our knowledge, to (Duboue and McKeown, 2003). Furthermore, although we had initially planned to produce a training set for the task, the cost of manual annotation turned out to be prohibitive and the resulting corpus was only fit for development and baseline evaluation. Despite these setbacks, we believe that open semantic web data is a promising test-bed and application field for NLG-oriented content selection (Bouayad-Agha et al., 2013) and trust that this first challenge has prepared the ground for follow up challenges with a larger participation. We would also like to encourage researchers from NLG and Semantic Web research fields to exploit the framework and materials developed during the course of this challenge to advance research in content selection.

## References


Nadjet Bouayad-Agha, Gerard Casamayor, and Leo Wanner. 2013. Natural Language Generation in the Context of the Semantic Web. *Submitted to the Semantic Web Journal*.

Nadjet Bouayad-Agha, Gerard Casamayor, Chris Mellish, and Leo Wanner. 2012. Content Selection from Semantic Web Data. *INLG '12 Proceedings of the Seventh International Natural Language Generation Conference*. Pages 146-149.

Pablo A. Duboue and Kathleen R. McKeown. 2003. Statistical Acquisition of Content Selection Rules for Natural Language Generation *Proceedings of the Conference on Empirical Methods for Natural Language Processing (EMNLP)*. Pages 121–128.

Randolph, Justus J. 2005. Free-marginal multirater kappa (multirater Kfree): An alternative to fleiss fixed-marginal multirater kappa. *Presented as the Joensuu University Learning and Instruction Symposium*.

Roman Kutlak, Chris Mellish and Kees van Deemter 2013. Content Selection Challenge University of Aberdeen entry *Proceedings of the 14th European Natural Language Generation (ENLG) Workshop*.

Hareen Venigalla and Barbara Di Eugenio. 2013. UIC-CSC: The Content Selection Challenge Entry from the University of Illinois at Chicago *Proceedings of the 14th European Natural Language Generation (ENLG) Workshop*.