Leveraging NLP Technologies and Linked Open Data to Create Better CAT Tools

Chris Hokamp
CNGL Centre for Global Intelligent Content
Dublin City University, School of Computing
Dublin, Ireland
chokamp@computing.dcu.ie

Abstract

This paper presents a prototype of a Computer Aided Translation (CAT) interface integrated with an entity extraction system to create a dynamic linked terminology component. The entity extraction system tags terms in the source sentence, mapping them to translation candidates in the target language. A usage scenario for linked data within a CAT tool is evaluated by prototyping all components necessary to construct a real-time dynamic terminology. By making use of Natural Language Processing (NLP) technologies including entity linking (Mihalcea *et al.* 2007), and statistical models for extracting and disambiguating entities (Daiber *et al.* 2013), the tool can provide translators with rich feedback about potential target-language translations of entities in the source text.

Keywords: *Post-editing, Machine Translation, CAT, Linked Open Data*

1. Introduction

Linked Open Data (LOD) can potentially be utilized at many points in the localisation workflow. By augmenting the metadata for the source or target text in a pre- or post-processing phase, linked data can provide metadata which facilitates human translation and quality assessment. Where metadata can be added in a completely automatic way, the task of determining whether or not the data is useful in the context can be pushed to the translator, who can decide where and how to make use of the additional information. The feedback from translators can then be used to augment the knowledge base.

In the dynamic terminology component presented here, a LOD resource and a statistical entity linker are combined within a translator-in-the-loop system. Translator-in-the-loop means that the design of the system explicitly includes a human, who is finally responsible for selecting the correct translation. This setup can be contrasted with a fully automatic design, where the target sentence would automatically be augmented with terminology, either via a machine translation system, or via an automatic post-editing phase.

The terms **entity** and **surface form** as used in this paper are defined as follows: an **entity** is concept (usually noun-like), represented by a unique DBPedia URI (Lehmann *et al.* 2014). A **surface form** is the text that is used to link to that entity – in other words, it is a language-specific string used to describe the entity. In Wikipedia, surface forms appear as blue hyperlinks in text (this indicates that an editor has linked the text with another page in Wikipedia). The set of possible surface forms for a given entity can be created by aggregating all links to the entity across all of a language's Wikipedia version, resulting in a (typically large) set of possible ways to refer to the entity. Table 1 shows the top ten German surface forms for the DBPedia entity "Earth".

DBPedia URI: http://dbpedia.org/resource/Earth
Surface Forms
Erde
Erdoberfläche
Welt
Erdbahn
Erdkugel
Terra
Planet
Erdkörpers
irdischen
Terra

Table 1: The most frequent German surface forms for the DBPedia entity "Earth"

In our design, the linking system tagger detects entities in a source segment, and the LOD resource provides candidate translations in the target language. By leveraging Wikipedia's multilingual graph through the DBPedia datasets, the system can provide suggestions for many language pairs. The multilingual graph of entities is thus transformed into a *dynamic terminology database*.

The term *dynamic* in this context means that the set of suggestions for a term depend upon the context in which it is being used. Because the disambiguation is done with respect to the context, the possible target forms are ranked according to their likelihood with respect to the underlying entity. A central hypothesis of this work is that this dynamic re-ranking provides a major improvement over the standard glossary or terminology lookup, which can only look for string matches for a particular token or phrase, without regard to the particular sense of the term in context.

2. Related Work

The majority of work on linked data for translation has focused on creating standards for data exchange, and on connecting backend resources such as terminologies with LOD ontologies. However, integrating linked data into the localisation workflow is an active area of research, and several projects, notably the ongoing FALCON (Lewis 2014b) project, are evaluating potential usecases as part of their goal to develop standards for linked data in localisation workflows. A prototype web-based application which integrates metadata using the Internationalization Tag Set (ITS) (Filip *et al.* 2013) within XLIFF 1.2 (Savourel *et al.* 2008) documents is presented in Porto *et al.* (2013).

3. Component Design

The motivating hypothesis for the component design is that the most difficult part of translating terminology is selecting the correct surface form for an entity. In other words, determining which entity a source language string refers to is easier than determining the correct translation for an entity, because of the nuance involved in choosing the correct surface form in the target language (formal vs. colloquial, full name vs. abbreviation, etc...). Thus, the component does not attempt to automatically select the correct surface form. The target side component is pre-populated with the translation options (ranked by frequency), and the translator must select the best option from the candidates. This design is similar to a terminology lookup or translation memory UI component, presenting translators with options instead of automatically populating the target translation with a "best" hypothesis.

We make use of the DBPedia-Spotlight (Daiber *et al.* 2013) statistical backend to perform the entity extraction step. The multilingual links in Wikipedia allow a mapping between languages to be created, so that concepts in the source language can be connected with concepts in the target language (the language specific URIs for a concept point to the same unique DBPedia URI). After the source entities have been extracted, target language translation candidates are found by moving in the opposite direction, generating the possible set of surface forms from the entity. The target-side surface forms are ranked by occurrence count with respect to the entity. Figure 1 shows a simple schematic of the flow of data through the component.

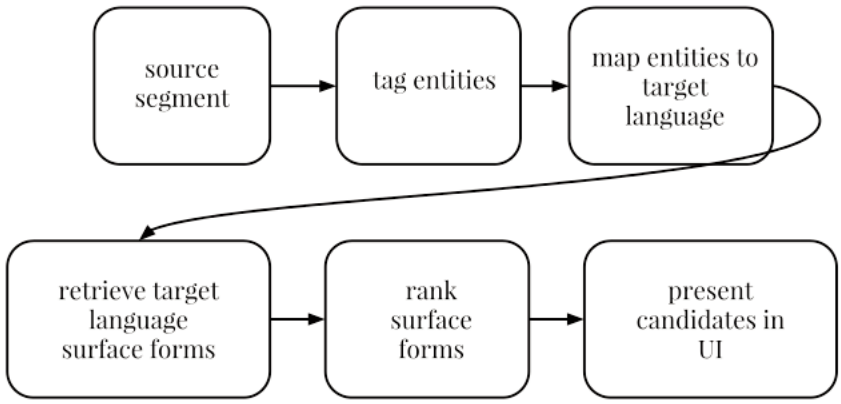

Figure 1: The dynamic linked terminology workflow

4. HandyCAT

HandyCAT is a flexible web based CAT tool (Lewis *et al.* 2014), specifically designed with interoperability and extensibility in mind. Because graphical components can easily be added and removed from the interface, it is an ideal platform for developing prototypes. The dynamic linked terminology component is designed as a standalone module that can easily added or removed from HandyCAT.

The server components are designed as microservices which are accessed using RESTful APIs, each fulfilling a single task in the dynamic terminology building process. The system is designed to operate in realtime, meaning that it does not require any offline preprocessing of the translation job.

5. Rendering Translation Options

Figure 2 shows a screenshot from an actual post-editing session. The user is evaluating the translation options for the source word "Europe". Upon selecting the best option (in this case the first option), the term will be inserted into the target area on the right side. All of the terms and markup are determined *automatically* and *on-the-fly* by the system, that is, not interfere with the normal operation of the interface.

6. Entity Linking and Labeling

Resources such as DBpedia and Freebase (Bollacker *et al.* 2008) are examples of open knowledge bases which take advantage of the implicit and explicit links in Wikipedia and other resources to construct a graph of entities with edges encoding relationships between the entities.

The process of finding the target-language surface form for a source entity requires two disambiguation steps. The first step is *entity linking*, where the entity extraction system attempts to link surface forms in the source language to the specific entity they represents. See Daiber *et al.* (2013) for details on the algorithm used to determine which entity is most likely represented given a surface form and a surrounding context.

The second step is *entity labeling*, where the translator selects the correct surface form for the entity in the target language. This requires retrieving the set of target language links for each entity, and making them available in the translation interface.

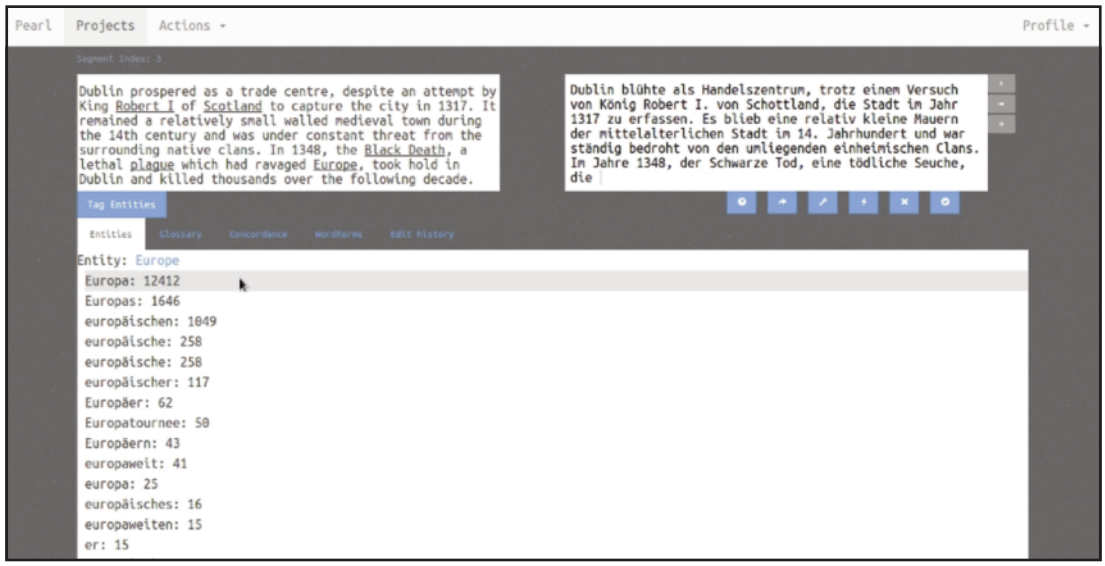

Figure 2: A screenshot of a HandyCAT editing session

there is no hard-coding of any entities or surface forms in either language, and the source text is parsed and linked when the translator enters the segment. The system can generate translation options for the entities in a source sentence in less than one second, so it does

6.1 Limitations

The terminology currently available to the component is limited to the data contained in Wikipedia, a resource that would probably not have good coverage

for many translation tasks, especially in specialized domains. Furthermore, the accuracy of the system is dependent upon the accuracy of the extraction framework. If a source entity is linked incorrectly, the translator could be presented with incorrect translation options.

Although the performance of the entity linking system is quite good, it is not perfect, so deploying the tool as part of a localisation workflow would require translators to carefully audit the translation options to ensure that they are not being presented with options that are generated from an incorrect entity. The entity linking component also requires many training examples for each entity in order to achieve good accuracy, so adding entities not contained in a large open dataset would necessitate curating a new training dataset – a process which could turn out to be prohibitively time-consuming.

7. Future Work

Evaluation of user interface components must be conducted with respect to a metric that can be measured in controlled user tests with and without the component present in the interface configuration. Some possible evaluation metrics are listed in table 2. Formal evaluation with one or more of these metrics has not yet been conducted, and the current prototype is simply a proof-of-concept.

Because the component is factored into standalone backend services (entity extraction and surface form mapping) and user interface elements, it can serve as simple enhancement to an existing interface. The backend services could also be integrated into a Machine Translation (MT) system, so that entities are added to the translation options considered by the MT system, instead of explicitly asking the user to choose the correct surface forms for the source entities.

8. Potential Integration with XLIFF and ITS

The tagging and disambiguation frameworks presented in this paper could be used as a standalone components in any localisation workflow. ITS and XLIFF are ideal for persisting translation options, and translators' choices for the best candidates in a particular context (Porto *et al.* 2013). One potential usecase could be used to add terminology to a project before it is sent to translators, allowing the information to be downstream in the translation process.

9. Conclusion

There are many opportunities to integrate existing NLP technologies into the Computer Aided Translation pipeline, but very few functional prototypes have been created to date. This work presented an end-to-end prototype of a dynamic linked terminology component implemented as part of the HandyCAT platform. The component was created to demonstrate a potential usecase for linked data within the localisation workflow, and to evaluate the effort needed to build such a system. This system enhances the resources available to translators without forcefully guiding the translation process, because translators are free to completely ignore the additional markup and terminology options if they wish. We believe that the human-in-the-loop paradigm is ideal for many CAT components, because it allows

CAT Component Evaluation Metrics
translator speed (words/min, segments/hour, etc…)
keypresses/operations per segment
quality of the resulting translation (human or automatic evaluation)
cognitive load (as measured by eye tracking or other methods)

Table 2: metrics for formally evaluating CAT UI components

translators to take advantage of additional metadata without requiring them to utilize the component(s) in cases where they do not perceive additional value.

Acknowledgments

This work was supported by the European Commission FP7 EXPERT project.

References

Bollacker, K., Evans, C., Paritosh, P., Sturge, T., Taylor, J. (2008) 'Freebase: A collaboratively created graph database for structuring human knowledge', in *Proceedings of the 2008 ACM SIGMOD Interna-tional Conference on Management of Data,* New York, NY, USA. ACM, 1247–1250.

Comerford, T., Filip, D., Raya, R.M., Savourel, Y. (Eds.) (2014) XLIFF Version 2.0 [online], OASIS Standard. ed, Standard, OASIS, available: http://docs.oasis-open.org/xliff/xliff-core/v2.0/os/xliff-core-v2.0-os.html [accessed 22 Aug 2014].

Daiber, J., Jakob, M., Hokamp, C., Mendes, P. (2013) 'Improving efficiency and accuracy in multilingual entity extraction', in *Proceedings of the 9th International Conference on Semantic Systems,* New York, NY, USA. ACM, 121–124.

Filip, D., McCance, S., Lewis, D., Lieske, C., Lommel, A., Kosek, L., Sasaki, F., Savourel, Y. (Eds.) (2013) Internationalization Tag Set (ITS) Version 2.0 [online], W3C Recommendation, W3C, available: http://www.w3.org/TR/its20/ [accessed 16 May 2014].

Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P., Hellmann, S., Morsey, M., van Kleef, P., Auer, S., Bizer, C. (2014) 'DBpedia – A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia', *Semantic Web Journal, 6(2), 167-195.*

Lewis, D., Liu, Q., Finn, L., Hokamp, C., Sasaki, F., Filip, D. (2014a). 'Open, Web-based Interna-tionalization and Localization Tools', *Translation Spaces*, vol III.

Lewis, D. (2014b) FALCON [online], available: http://falcon-project.eu/wp-content/uploads/2014/05/FALCON-Poster-mlw-madrid-may141.pdf [accessed 15 May 2014].

Mihalcea, R., Csomai, A. (2007). 'Wikify!: Linking Documents to Encyclopedic Knowledge', in P*roceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, New York, NY, USA. ACM, 233–242.

Porto, P., Lewis, D., Finn, L., Saam, C., Moran, J., Serikov, A., O'Connor, A. (2013). 'ITS2.0 and Computer Assisted Translation Tools'. *Localisation Focus - The International Journal of Localisation*, 12.

Savourel, Y., Reid, J., Jewtushenko, T., Raya, R. (Eds.) (2008) XLIFF Version 1.2 [online], OASIS Standard. ed, Standard, OASIS, available: http://docs.oasis-open.org/xliff/v1.2/os/xliff-core.html [accessed 15 May 2014].