EUROPEAN COOPERATION
IN SCIENCE & TECHNOLOGY

WEB PORTAL DEVELOPMENT
STSM REPORT

Action number: IS1312
STSM title: Web Portal Development
STSM start and end date: 06/11/2017 to 17/12/2017
COST Applicant: Murathan Kurfalı, Middle East Technical University
Host: Prof. Dr. Bonnie Webber, University of Edinburgh

PURPOSE OF THE STSM:

The main objective behind the "Web Portal Development" STSMs is to prepare the Textlink Web Portal which will ultimately provide a various search, visualization and dissemination capabilities for researchers. The current STSM constituted the second phase and specifically aimed to develop the existing portal further by giving access to aligned TED-Multilingual Discourse Bank(TED-MDB) annotations as well as providing access to dictionary of connectives available in DimLex format.

Firstly, the portal aims to enable multi-lingual linguistic investigation on discourse level through aligned tokens. Through the web portal users will be able to access the annotations in any language available (German, Russian, Polish, Portuguese, Turkish) and their counterparts in English. That is, STSM aims to enable researchers to see how discourse relations covering the same discourse units are realized in different languages.

Secondly, in order to extend the usability of the web portal with more detailed linguistic information such as syntactic structure, part-of- speech tags, dependency parses, word senses or DRD specific information, readily available DimLex lexicons are linked to the web-portal.

DESCRIPTION OF WORK CARRIED OUT DURING THE STSM

Firstly, in order to build the web site, an Ubuntu Server 16.04 was launched in a virtual instance at Amazon Web Services (AWS). Then, in order to resolve the problems that have been encountered during the first STSM, the currently available facilities of the web-site is re-implemented using the Jingo Framework(Version 1.11.6, http://www.djangoproject.com/) (The first version was implemented using Drupal 8). In order to implement the desired functionalities, Python, MYSQL, JavaScript (and JQuery), AJAX and HTML are used.

In order to display information of the existing DimLex lexicons, a dynamic link between connective-lex.info and the web portal is created. Therefore, the portal is able to fetch information from the connective-lex.info on-demand in order to ensure the consistency between two web-sites.

Funded by the Horizon 2020 Framework Programme

As for, multilingual part, aligned tokens in two TED-MDB talks are uploaded the portal's database. Currently, the annotations are aligned between language pairs X-ENG (where X is German, Russian, Polish, Portuguese, Turkish).

Specifically via the multi-lingual mode, users can

- Access the aligned tokens of TED-MDB directly as the annotations are uploaded the databases permanently.
- Filter the annotations in any language using the the search options which are also available for monolingual mode (basically, users can filter the annotations according to its *sense* (and/or) *connective* (and/or) *type*).
- Retrieve all corresponding relations in the target language.
- Automatically highlight the corresponding annotation in the target language when a token is selected in the source language.

Thanks to the DimLex mode, users can

- access the senses conveyed by the selected connective
- access the frequencies regarding how many times each sense is conveyed by the selected connective
- (If available in DimLex lexicon for that language), access the meta-data such as part-of- speech tags, orthographic variants of the selected connective

Additionaly, appendix contains detailed desciprtion of the up-to-date situation of the web portal which is the outcome of the current STSM.

CONCLUSION

In summary, during the STSM, the existing web-portal is re-implemented using a more flexible framework (Django). Then, portal is enriched with multi-lingual mode which enables cross linguistics investigation on discourse level as well as with a DimLex module which provides detailed information regarding the discourse connectives.

TextLink Web Portal

Murathan Kurfalı[1], Ahmet Üstün[1], and Bonnie Webber[2]

[1]Informatics Institute, Middle East Technical University (ODTÜ)
{kurfali,ustun.ahmet}@metu.edu.tr
[2]School of Informatics, University of Edinburgh, bonnie@inf.ed.ac.uk

January 8, 2018

1 Introduction

This paper introduces the TextLink Web Portal - an online web service for searching discourse-annotated text, that stands as one of the promised outcomes of the TextLink Cost action[1]. As implemented, the portal serves two purposes: (1) to enable researchers to display and filter discourse annotations according various parameters[2] and then, if desired, download the resulting set; (2) to provide access to the growing multi-lingual TED-MDB corpora, allowing researchers to examine cross-lingual parallel discourse annotation. As a web service, the TextLink Web portal requires no installation and is easy to master, yet it is capable of performing complex queries. The rest of the paper explains its capabilities in detail.

2 TextLink Web Portal

The TextLink Web Portal is being developed as a publicly available web-site for examining, annotating and summarizing discourse annotation in either monolingual text or parallel cross-lingual bi-texts.

Currently, the portal has four sections:

- *Home*: The home page contains links to other pages of the web portal as well as to the main web site of TextLink. Users can also download sample files and the user manual from the home page.
- *Upload Annotations*: This page allows users to upload text files and their (stand-off) discourse annotation, on which they want to perform searches. Users can access this page any time they want to upload additional files. When uploading files, users should indicate the language of the annotations so that, whenever necessary, portal can retrieve information from the relevant DIMLex corpora.
- *Search (Monolingual)*: This page allows users to display their annotations and filter them through various search options. Section 2.1 gives a detailed description of the search page.
- *TED-MDB Search (Multilingual)*: This page hosts aligned annotation of files from the TED-MDB, to allow researchers to make cross-lingual comparison in discourse level annotations among the covered languages (see Section 2.2).

In the near future, we plan to implement an online annotator, similar to the PDTB annotator[3], where users can upload new text files and annotate their discourse relations without having to have a local copy of the PDTB annotator. In addition to existing abilities of PDTB annotator, the online annotator will be able to provide information regarding annotated tokens based on the previous annotations and (where available) DimLex lexicons.

[1] www.textlink.ii.metu.edu.tr

[2] Currently, portal only accepts annotations produced by either PDTB Annotator(Lee et al., 2016) or DATT(Aktaş, Bozsahin, & Zeyrek, 2010)

[3] http://www.seas.upenn.edu/ pdtb/annotator.html

2.1 Monolingual Search

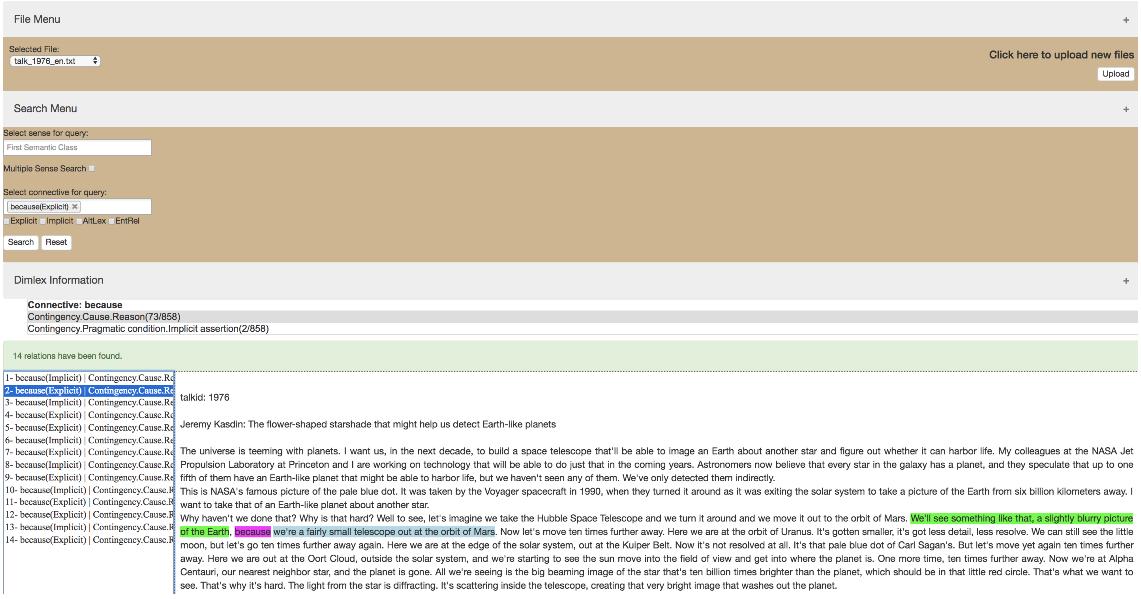
2.1.1 User Interface

The user interface consists of the following three main blocks (Figure 1):

- *Search panel*: The search panel resides on the top of the page. It allows a user to select annotation files, determine the search parameters and, provided that a connective is selected, display the list of the senses conveyed by the selected connective using connective-lex.

- *Annotation list*: The list on the left-hand side presents all annotation tokens for the file being searched. The list is updated when a user selects another file or performs a search. The selected annotation token becomes highlighted on the text. In the list, each annotation is represented with the discourse connective, if any, along with the type of the relation and the senses, if any, it conveys.

- *Text Panel*: The main panel displays the text file which was annotated. When an annotation is selected, it automatically scrolls to the annotation.



Selected File: talk_1976_en.txt

Select sense for query: First Semantic Class

Multiple Sense Search □

Select connective for query: because(Explicit) ✕

□ Explicit □ Implicit □ AltLex □ EntRel

Connective: because
Contingency.Cause.Reason(73/858)
Contingency.Pragmatic condition.Implicit assertion(2/858)

14 relations have been found.

1- because(Implicit) | Contingency.Cause.Re
2- because(Explicit) | Contingency.Cause.Re
3- because(Implicit) | Contingency.Cause.Re
4- because(Explicit) | Contingency.Cause.Re
5- because(Explicit) | Contingency.Cause.Re
6- because(Implicit) | Contingency.Cause.Re
7- because(Explicit) | Contingency.Cause.Re
8- because(Implicit) | Contingency.Cause.Re
9- because(Explicit) | Contingency.Cause.Re
10- because(Implicit) | Contingency.Cause.R
11- because(Explicit) | Contingency.Cause.R
12- because(Explicit) | Contingency.Cause.R
13- because(Implicit) | Contingency.Cause.R
14- because(Explicit) | Contingency.Cause.R

talkid: 1976

Jeremy Kasdin: The flower-shaped starshade that might help us detect Earth-like planets

The universe is teeming with planets. I want us, in the next decade, to build a space telescope that'll be able to image an Earth about another star and figure out whether it can harbor life. My colleagues at the NASA Jet Propulsion Laboratory at Princeton and I are working on technology that will be able to do just that in the coming years. Astronomers now believe that every star in the galaxy has a planet, and they speculate that up to one fifth of them have an Earth-like planet that might be able to harbor life, but we haven't seen any of them. We've only detected them indirectly.
This is NASA's famous picture of the pale blue dot. It was taken by the Voyager spacecraft in 1990, when they turned it around as it was exiting the solar system to take a picture of the Earth from six billion kilometers away. I want to take that of an Earth-like planet about another star.
Why haven't we done that? Why is that hard? Well to see, let's imagine we take the Hubble Space Telescope and we turn it around and we move it out to the orbit of Mars. We'll see something like that, a slightly blurry picture of the Earth, because we're a fairly small telescope out at the orbit of Mars. Now let's move ten times further away. Here we are at the orbit of Uranus. It's gotten smaller, it's got less detail, less resolve. We can still see the little moon, but let's go ten times further away again. Here we are at the edge of the solar system, out at the Kuiper Belt. Now it's not resolved at all. It's that pale blue dot of Carl Sagan's. But let's move yet again ten times further away. Here we are out at the Oort Cloud, outside the solar system, and we're starting to see the sun move into the field of view and get into where the planet is. One more time, ten times further away. Now we're at Alpha Centauri, our nearest neighbor star, and the planet is gone. All we're seeing is the big beaming image of the star that's ten billion times brighter than the planet, which should be in that little red circle. That's what we want to see. That's why it's hard. The light from the star is diffracting. It's scattering inside the telescope, creating that very bright image that washes out the planet.

Figure 1: A screenshot of the TextLink Web Portal search page

2.1.2 Search Facilities

The portal offers various filtering options. The search can be performed via interface without needing to write any SQL queries. The search options are listed below:

- *Sense Search*: One can search for tokens with a particular sense or senses by selecting them from the drop-down menu or typing or typing them in. Users can select as many senses as they want. The selected senses are combined by *or*, meaning that all relations which involve at least one of the selected senses will be included in the result set.

  Furthermore, as it is not uncommon for a discourse relation to convey two senses simultaneously, users can specify two senses, as well as how they should be combined using the operator menu. There are two possible operator, ‘and’– ‘not’. When the latter is selected, all the relations conveying the first sense but not the second one are retrieved.

- *Type Search*: Users can specify the type of the relation they want by selecting the check-boxes provided. Users can select as many check-boxes as they want.

- *Connective Search*: Finally, users can filter the annotations according to the discourse connective they possess. Users can select as many connective as they want by either typing or using the drop down list. Although PDTB anchors implicit discourse relations to a connective, which is referred as ‘implicit connective ’, the connective search is limited to Explicit and AltLex relations, as insertion of implicit connectives is prone to inconsistency.

All search parameters can be combined. That is, users can perform searches of the following kind: retrieve all ‘*Explicit*’ relations which convey an ‘*Expansion*’ sense but ‘*not*’ any of the ‘*Temporal*’ senses via the connective ‘*and*’.

For the time being, although users can upload as many files as they want, search can be performed on only one file at a time. However, we plan to extend this so that the specified search criteria will be applied to all the files uploaded.

Finally, the annotations uploaded by users are only stored in the portal during their session. That is, the portal does not store any files permanently, in order to provide confidentiality to users. Therefore, anyone can use the portal without risking their annotations to be publicized without their permission.

2.1.3 Download Facility

Portal enables users to download their search results for further processing. The results can be saved in the original file format (e.g. pipe delimited file) or as a CSV file. The difference is, CSV file contains text spans rather than byte spans (or XML tags, as in the case of DATT), rendering the results more readable, as well as suitable to process with office tools such as Excel. On the other hand, anyone who wishes to further process the results through the portal (or the Annotator where the annotations are prepared) can save them in the original pipe-delimited format.

2.1.4 DIMLex Facility

The TextLink Web Portal also incorporates available DIMLex-style lexicons through the connective-lex web-site[4]. Briefly, connective-lex provides an interface to perform search on available DIMLex corpora. Among others, connective-lex provides the list of the senses conveyed by any given connective. In the portal, when a user performs a search containing a connective , portal displays that list on the search page enabling to compare between the annotations included in the result set and all possible senses conveyed by the given connective according to DIMLex. Portal periodically retrieves DIMLex files from connective-lex through its API in order to avoid any inconsistency between two sites.

2.2 Multilingual Search (TED-MDB)

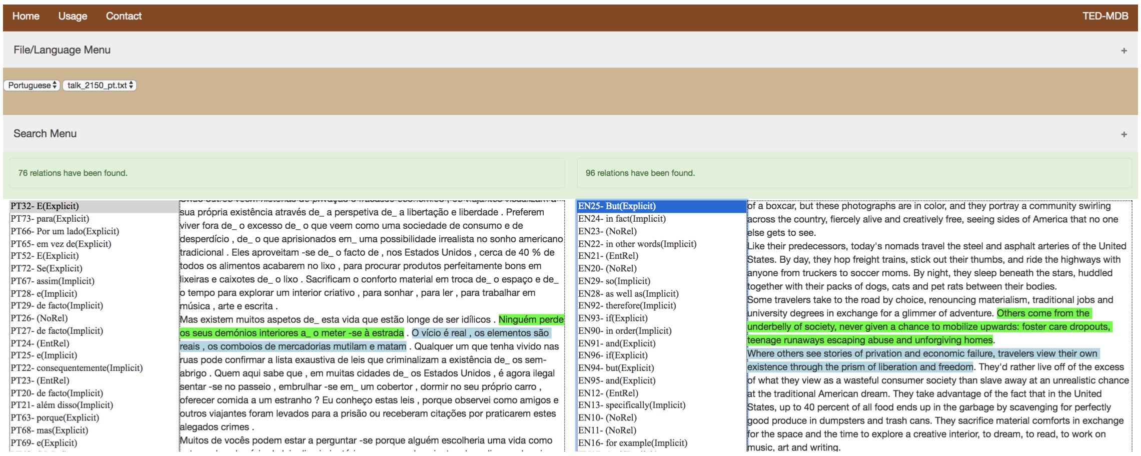
TED-MDB is a multilingual corpus of TED-talks, annotated in the style of the PDTB, currently covering six languages (English, European Portuguese, German, Polish, Russian, Turkish). Recently, the discourse relations of two talks in each language have been aligned with respect to English through semi-automatic means. Multilingual search option of the portal enables to access aligned tokens of the TED-MDB corpus.

A sample view of the page is provided in Figure 2. The search options are the same as those of monolingual search; however, there are two text panels where the rightmost one always displays the English tokens. Users can select among the languages and the files using the drop-down menus in the File/Language menu. Only the tokens in the selected language are filtered, as it is likely that aligned tokens convey different senses or may be of different types across languages. Whenever, a token is selected on the left annotation list, its English counterpart, if any, is automatically selected and highlighted.

3 Implementation Details

The portal currently resides on a Amazon EC2 server. The server-side logic is implemented using Django Framework(Version 1.11.6, http://www.djangoproject.com/) making use of Structured Query Language (SQL) in order to retrieve the desired annotation tokens from the databases. All queries are generated automatically from the user’ selections via interface. The interfaces, which are referred as *templates* in Django, are coded in HyperText Markup Language (HTML) with JavaScript, including AJAX, in order to provide necessary functionality.

[4] http://connective-lex.info/



76 relations have been found.

PT32- E(Explicit)
PT73- para(Explicit)
PT66- Por um lado(Explicit)
PT65- em vez de(Explicit)
PT52- E(Explicit)
PT72- Se(Explicit)
PT67- assim(Implicit)
PT28- e(Implicit)
PT29- de facto(Implicit)
PT26- (NoRel)
PT27- de facto(Implicit)
PT24- (EntRel)
PT25- e(Implicit)
PT22- consequentemente(Implicit)
PT23- (EntRel)
PT20- de facto(Implicit)
PT21- além disso(Implicit)
PT63- porque(Explicit)
PT68- mas(Explicit)
PT69- e(Explicit)

sua própria existência através de_ a perspetiva de_ a libertação e liberdade . Preferem viver fora de_ o excesso de_ o que veem como uma sociedade de consumo e de desperdício , de_ o que aprisionados em_ uma possibilidade irrealista no sonho americano tradicional . Eles aproveitam -se de_ o facto de , nos Estados Unidos , cerca de 40 % de todos os alimentos acabarem no lixo , para procurar produtos perfeitamente bons em lixeiras e caixotes de_ o lixo . Sacrificam o conforto material em troca de_ o espaço e de_ o tempo para explorar um interior criativo , para sonhar , para ler , para trabalhar em música , arte e escrita .

Mas existem muitos aspetos de_ esta vida que estão longe de ser idílicos . Ninguém perde os seus demónios interiores a_ o meter -se à estrada . O vício é real , os elementos são reais , os comboios de mercadorias mutilam e matam . Qualquer um que tenha vivido nas ruas pode confirmar a lista exaustiva de leis que criminalizam a existência de_ os sem-abrigo . Quem aqui sabe que , em muitas cidades de_ os Estados Unidos , é agora ilegal sentar -se no passeio , embrulhar -se em_ um cobertor , dormir no seu próprio carro , oferecer comida a um estranho ? Eu conheço estas leis , porque observei como amigos e outros viajantes foram levados para a prisão ou receberam citações por praticarem estes alegados crimes .

Muitos de vocês podem estar a perguntar -se porque alguém escolheria uma vida como

96 relations have been found.

EN25- But(Explicit)
EN24- in fact(Implicit)
EN23- (NoRel)
EN22- in other words(Implicit)
EN21- (EntRel)
EN20- (NoRel)
EN29- so(Implicit)
EN28- as well as(Implicit)
EN92- therefore(Implicit)
EN93- if(Explicit)
EN90- in order(Implicit)
EN91- and(Explicit)
EN96- if(Explicit)
EN94- but(Explicit)
EN95- and(Explicit)
EN12- (EntRel)
EN13- specifically(Implicit)
EN10- (NoRel)
EN11- (NoRel)
EN16- for example(Implicit)

of a boxcar, but these photographs are in color, and they portray a community swirling across the country, fiercely alive and creatively free, seeing sides of America that no one else gets to see.

Like their predecessors, today's nomads travel the steel and asphalt arteries of the United States. By day, they hop freight trains, stick out their thumbs, and ride the highways with anyone from truckers to soccer moms. By night, they sleep beneath the stars, huddled together with their packs of dogs, cats and pet rats between their bodies.

Some travelers take to the road by choice, renouncing materialism, traditional jobs and university degrees in exchange for a glimmer of adventure. Others come from the underbelly of society, never given a chance to mobilize upwards: foster care dropouts, teenage runaways escaping abuse and unforgiving homes.

Where others see stories of privation and economic failure, travelers view their own existence through the prism of liberation and freedom. They'd rather live off of the excess of what they view as a wasteful consumer society than slave away at an unrealistic chance at the traditional American dream. They take advantage of the fact that in the United States, up to 40 percent of all food ends up in the garbage by scavenging for perfectly good produce in dumpsters and trash cans. They sacrifice material comforts in exchange for the space and the time to explore a creative interior, to dream, to read, to work on music, art and writing.

Figure 2: A screenshot of the TextLink Web Portal TED-MDB search page

References

Aktaş, B., Bozsahin, C., & Zeyrek, D. (2010). Discourse relation configurations in turkish and an annotation environment. In *Proceedings of the fourth linguistic annotation workshop* (pp. 202–206).

Lee, A., Prasad, R., Webber, B. L., & Joshi, A. K. (2016). Annotating discourse relations with the pdtb annotator. In *Coling (demos)* (pp. 121–125).