Documento di Progetto
SMACH_R_WP3_t1_YIS2
Report SMACH-WP3, anno 2, semestre 1
03/01/2021

SMACH
Semantic Multi-lingual Access to Cultural Heritage

PON "Ricerca e Innovazione" 2014/2020
Asse I "Capitale Umano"
Azione I.2 "Attrazione e Mobilità dei Ricercatori"

WP 3 - Task 1
"Il Corpus CHerIDesCo (Cultural Heritage - Italian Description Corpus)"

Gloria Gagliardi
Massimo Guarino

Unior NLP research Group
Dipartimento di Studi Letterari, Linguistici e Comparati
Università degli Studi di Napoli "L'Orientale"

Indice

1. Il corpus CHerIDesCo (Cultural Heritage - Italian Description Corpus): presentazione della risorsa

1.1 Introduzione

La costruzione del Corpus CHerIDesCo (Cultural Heritage - Italian Description Corpus) si inserisce nelle attività di ricerca previste nel task 1 del WP3 del progetto SMACH.

In lingua italiana, infatti, la possibilità di applicare approcci TAL ai BBCC è fortemente limitata dall'assenza di corpora di dominio, monolingui e multilingui, da utilizzare per il *training* e il *testing* degli algorimi.

CHerIDesCo si propone di fornire una prima soluzione a tale problema: è infatti una raccolta bilanciata [1, 2] di testi descrittivi prodotti dalle istituzioni riferiti ai musei, monumenti e siti archeologici statali italiani.

1.2 Descrizione della Risorsa

CHerIDesCo è un corpus sincronico monolingue che raccoglie testi scritti descrittivi in lingua italiana riferiti a musei, monumenti e siti archeologici statali.

I testi sono stati raccolti a partire dagli elenchi dei musei statali messi a disposizione dal MiBACT (`http://musei.beniculturali.it/musei`).

Il corpus e le relative annotazioni, disponibili all'URL `https://github.com/unior-nlp-research-group/SMACH-corpora`, sono completamente aperti e utilizzabili senza restrizioni per scopi di ricerca.

1.2.1 Dimensione e struttura del corpus

La risorsa si compone, nella sua versione 1.0, di 680 testi (estensione: *.txt) di varia lunghezza,[1] per un totale di 233016 parole, organizzati in 17 subcorpora regionali:[2]

- CHerIDesCo–Abruzzo_subcorpus: 23 siti censiti, 20570 parole;
- CHerIDesCo–Basilicata_subcorpus: 21 siti censiti, 5063 parole;
- CHerIDesCo–Calabria_subcorpus: 25 siti censiti, 13943 parole;
- CHerIDesCo–Campania_subcorpus: 103 siti censiti, 38433 parole;
- CHerIDesCo–EmiliaRomagna_subcorpus: 42 siti censiti, 13312 parole;
- CHerIDesCo–FriuliVeneziaGiulia_subcorpus: 18 siti censiti, 5041 parole;

[1] Da un minimo di 7 a un massimo di 5070 parole; media = 343.

[2] Sono esclusi dal corpus i Siti delle Regioni a Statuto Speciale Valle d'Aosta, Trentino Alto Adige e Sicilia, che gestiscono e coordinano autonomamente musei, aree e parchi archeologici e monumenti del proprio territorio.

- CHerIDesCo–Lazio_subcorpus: 135 siti censiti, 46275 parole;
- CHerIDesCo–Liguria_subcorpus: 15 siti censiti, 3534 parole;
- CHerIDesCo–Lombardia_subcorpus: 26 siti censiti, 7588 parole;
- CHerIDesCo–Marche_subcorpus: 31 siti censiti, 7238 parole;
- CHerIDesCo–Molise_subcorpus: 14 siti censiti, 8688 parole;
- CHerIDesCo–Piemonte_subcorpus: 22 siti censiti, 6936 parole;
- CHerIDesCo–Puglia_subcorpus: 21 siti censiti, 7628 parole;
- CHerIDesCo–Sardegna_subcorpus: 60 siti censiti, 14469 parole;
- CHerIDesCo–Toscana_subcorpus: 72 siti censiti, 24840 parole;
- CHerIDesCo–Umbria_subcorpus: 14 siti censiti, 4938 parole;
- CHerIDesCo–Veneto_subcorpus: 19 siti censiti, 4520 parole.

1.2.2 Metadati

A ciascun oggetto della risorsa (i.e. corpus, subcorpus, testo) è associato un file di metadati in formato xml (estensione: *.imdi), compilato seguendo le definizioni previste dallo standard internazionale IMDI (ISLE Meta Data Initiative). Per la creazione e la gestione dei metadati è stato utilizzato il software Arbil [5].[3]

1.2.3 Annotazione

L'annotazione del corpus CHerIDesCo è stata effettuata ricorrendo ai diversi tool contenuti in *Stanza*[4], un ***python package*** di strumenti di Natural Language Processing predisposto dallo Stanford NLP Group. La pipeline utilizzata al fine dell'annotazione del corpus contiene i processori *token* (che divide il testo in frasi e, successivamente al loro interno, tokens), *Part of Speech* (POS - che genera le annotazioni *UPOS*, *XPOS* e *Feats*) e, infine, il processore *Lemma* che provvede alla lemmatizzazione delle parole utilizzando come input i singoli token ed i valori derivanti dall'annotazione della parte del discorso. In particolare, in relazione alla parte del discorso, si riporta sia l'annotazione "universale" (UPOS)[4] che quella inerente alla specifica treebank utilizzata (XPOS). Per quanto riguarda lo Universal POS tagging lo schema di annotazione è basato sui tag universali delle parti del discorso di Google [3] e sui tagsets morfosintattici di Interset [6]. In aggiunta al pos-tagging sono fornite, per ogni token, ulteriori caratteristiche lessicali e grammaticali (Universal Feats) le cui categorie sono elencate sul sito `https://universaldependencies.org/u/feat/index.html`

1.2.4 Aggiornamento ed accrescimento del corpus

La risorsa è stato progettato per crescere nel tempo, aumentando i punti di raccolta ed il numero di testi riferiti ai siti censiti. Ad ogni testo è infatti associato un ID univoco, così strutturato:

[3] https://archive.mpi.nl/forums/t/arbil-information-manuals-download/1045

[4] Tale modalità di annotazione è ampiamente descritta sul sito `https://universaldepencies.org`: "*Universal Dependencies (UD) is a project that is developing cross-linguistically consistent treebank annotation for many languages, with the goal of facilitating multilingual parser development, cross-lingual learning, and parsing research from a language typology perspective.*"

Regione_IDSito_IDtesto
es. Abruzzo_001_01 : *Museo archeologico nazionale di Campli - presentazione sito MiBACT*
es. Abruzzo_001_02 : *Museo archeologico nazionale di Campli - Depliant*
es. Abruzzo_002_01 : *Chiesa di San Pietro ad Oratorium, Capestrano - presentazione sito MiBACT*

1.3 Criticità

1.4 Proposta di Lavoro

CHerIDesCo si presta sia ad una valida estensione del lavoro contenuto nel WP2. t3.Y1S2, relativamente al quale fornisce un più ampio corpus per l'estrazione di Entità Nominate (NER), sia ad un'analisi approfondita dei diversi topic in cui può essere immaginata l'offerta dei beni culturali statali (attraverso l'impiego delle tecniche di apprendimento automatico inerenti alla Topic Analysis). Più specificamente, l'ultimo punto dovrebbe mirare all'individuazione dei differenti aspetti tematici, qualora esistenti, derivanti dall'eterogenea natura dei soggetti pubblici coinvolti dal lato dell'offerta dei beni culturali. Una siffatta analisi contribuirebbe in modo più incisivo ad evidenziare i punti di forza più importanti dell'intero settore dell'offerta pubblica dei beni culturali, potendo eventualmente individuare, al contempo, diverse configurazioni delle funzioni svolte da macro-raggruppamenti omogenei di diverse istituzioni culturali pubbliche.

Bibliografia

[1] E. Cresti and A. Panunzi. *Introduzione ai corpora dell'italiano*. il Mulino, Bologna, 2013.

[2] T. McEnery and A. Wilson. *Corpus Linguistics*. Edinburgh University Press, Edinburgh, 1996.

[3] Slav Petrov, Dipanjan Das, and Ryan McDonald. A universal part-of-speech tagset. *Computing Research Repository - CORR*, 04 2011.

[4] Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. Stanza: A python natural language processing toolkit for many human languages, 2020.

[5] P. Withers. Metadata management with arbil. In *Proceedings of LREC 2012: 8th International Conference on Language Resources and Evaluation*, pages 72–75, Paris, France, 2012. European Language Resources Association (ELRA).

[6] Daniel Zeman. Reusable tagset conversion using tagset drivers. In *LREC*, volume 2008, pages 28–30, 2008.