Versione PROVVISORIA del contributo presentato al Convegno Annuale

DNT: un Corpus Diacronico e Multigenere di Testi in Lingua Inglese

Tommaso Caselli[1], Rachele Sprugnoli[2]

[1] University of Groningen, The Netherlands, t.caselli(«»)rug.nl

[2] CIRCSE Research Centre, Università Cattolica del Sacro Cuore, Italia, rachele.sprugnoli(«»)unicatt.it

ABSTRACT

In questo contributo presentiamo il Diachronic News and Travel (DNT) corpus, una collezione di testi in lingua inglese che copre diversi generi testuali e due ampi periodi temporali. Più precisamente, il corpus è formato da testi sia contemporanei che storici di tre generi, ovvero articoli di quotidiano, resoconti di viaggio e guide turistiche. Il corpus, liberamente disponibile, è già stato usato per l'analisi di vari fenomeni linguistici e per lo sviluppo di sistemi per il Trattamento Automatico del Linguaggio che vengono qui brevemente descritti.

KEYWORDS

Corpus, Annotazione, Trattamento Automatico del Linguaggio.

1. INTRODUZIONE

Lo sviluppo e la diffusione di corpora e di risorse linguistiche ha visto una rapida crescita a partire dalla prima metà degli anni 2000. La *Language Resource and Evaluation Map*[1] riporta la documentazione di oltre 2,920 corpora in diverse lingue ma questa stima, basata su uno sforzo volontario della comunità, è destinata a crescere.

Sebbene la disponibilità di corpora sia stata favorita dal crescente sviluppo di tecnologie per il Trattamento Automatico del Linguaggio (TAL), un aspetto centrale dei corpora è quello di essere istanze di *performance* linguistiche [3]. La combinazione di disponibilità e uso di corpora ha introdotto un profondo cambiamento negli studi sul linguaggio, agevolando la verifica empirica di teorie e fenomeni linguistici [11][2][6].

Questo contributo introduce il Diachronic News and Travel (DNT) corpus, composto da testi scritti in lingua inglese. La sua caratteristica distintiva è la copertura sia di diversi generi testuali (i.e., articoli di quotidiano e scrittura di viaggio, quest'ultima a sua volta divisa tra resoconti in prima persona e guide turistiche) sia di diversi periodi temporali (i.e., 1860-1939 e 1998-2017).

2. DESCRIZIONE DEL CORPUS

La combinazione di periodi temporali e generi testuali diversi permette di dividere il corpus in 6 sotto-corpora le cui dimensioni in termini di numero di documenti e di token sono riportate nella Tabella 1.

Genere e Periodo Temporale	# Documenti	# Token
Articoli di quotidiano 1998-2017	84	32.086
Articoli di quotidiano 1860-1939	50	29.717
Resoconti di viaggio 1998-2017	23	30.747
Resoconti di viaggio 1860-1939	25	31.690
Guide turistiche 1998-2017	58	29.950
Guide turistiche 1860-1939	39	29.327
TOTALE	**279**	**183.517**

Tabella 1: Statistiche generali sul corpus

Nel creare la risorsa, si è cercato di bilanciare la dimensione dei vari sotto-corpora anche se un bilanciamento perfetto non è stato possibile a causa della difficoltà nel reperire sia testi storici puliti da eventuali errori dovuti all'OCR che testi contemporanei liberi da copyright. Per quanto riguarda questi ultimi, gli articoli di giornale sono stati estratti da altre risorse linguistiche (come il Wall Street Journal Corpus [4]), i resoconti di viaggio da post pubblicati online su blog di viaggiatori mentre i testi delle guide derivano dal portale WikiTravel[2] e dal sito della Lonely Planet[3]: per ciascun post e per le pagine della Lonely Planet abbiamo ottenuto il permesso di riuso. Relativamente ai testi storici, invece, gli articoli

[1] http://lremap.elra.info/?&selected_facets=resourceTypeFilter_exact%3ACorpus

[2] https://wikitravel.org/en/Main_Page

[3] https://www.lonelyplanet.com/

di giornale sono stati presi dal portale Wikisource[4] mentre resoconti di viaggio e guide sono stati scaricati dal sito del Progetto Gutenberg.[5]

3. CASI D'USO

Il DNT è stato usato prevalentemente per lo sviluppo di tecnologie di TAL. Questo ha di fatto imposto sullo stesso insieme di dati, o su un suo sottoinsieme, annotazioni manuali su diversi livelli, offrendo la possibilità di studiare l'interazione di fenomeni linguistici all'intersezione delle dimensioni temporali e/o di genere. Di seguito presentiamo brevemente tre casi di studio che mostrano come DNT possa essere utile per diverse analisi e applicazioni.

Il più complesso tra i casi di studio riguarda l'annotazione di micro-atti illocutori a livello di proposizioni realizzati tramite tipi di contenuto (o *content types*)[6] [12]. A differenza di lavori precedenti [5][10], un unico schema di annotazione è stato elaborato e applicato sull'intero corpus. L'analisi delle annotazioni ha evidenziato una distribuzione statisticamente significativa dei diversi tipi di contenuto nei diversi generi, ma non nei diversi periodi temporali. Al contrario, l'applicazione di un sistema supervisionato ha evidenziato come una certa differenza nella distribuzione dei tipi di contenuto sembra essere presente anche per la dimensione diacronica [8].

Gli altri due casi riguardano invece l'analisi dei soli testi storici. Lo studio della realizzazione linguistica degli eventi [1] ha visto l'integrazione di più schemi di annotazione in un unico modello e la sua applicazione ai vari generi presenti nel corpus. Questo ha evidenziato una distribuzione di classi di eventi diversi statisticamente significativa nelle scritture di viaggio e negli articoli di giornale. Un sistema supervisionato è poi stato addestrato sui dati annotati per il riconoscimento e la classificazione semantica degli eventi in testi storici[7] [7].

Infine, lo studio di entità nominate si è concentrato su guide e resoconti di viaggio.[8] L'applicazione di un sistema TAL allo stato dell'arte[9] ha messo in evidenza i limiti dell'applicazione di queste tecnologie a dati storici e/o generi diversi da quelli usati per sviluppare il sistema. In particolare, la presenza di fenomeni specifici quali variazioni ortografiche di nomi propri, frequenza di grafemi latini, uso di forme abbreviate ha richiesto l'annotazione manuale di nuovi testi e lo sviluppo di un nuovo sistema per il riconoscimento automatico di nomi propri di luoghi geografici (e.g., fiumi, monti), politici (e.g., città, nazioni) e funzionali (e.g., monumenti, siti archeologici) [9].

4. CONCLUSIONI

Nell'ottica di favorire e promuovere una cultura e una scienza aperta, la condivisione di dati e la replicabilità degli esperimenti, tutte le risorse descritte in questo contributo sono rilasciate liberamente. Nello specifico, il DNT corpus in formato testo è disponibile online[10] come lo sono anche le varie annotazioni prodotte nei casi d'uso sopra menzionati ed i modelli di Trattamento Automatico del Linguaggio sviluppati. Una versione *tokenizzata* del corpus sarà presto rilasciata per facilitare l'applicazione di altri livelli di annotazione da parte della comunità e promuovere lo sviluppo di un'ampia risorsa aperta e multi-livello.

BIBLIOGRAFIA

[1] Bach, Emmon. The algebra of events. Linguistics and philosophy (1986): 5-16.

[2] Chiari, Isabella and Jezek, Elisabetta. 2016. 'Dati Empirici e Risorse lessicali'. Introduzione al numero monografico RiCognizioni, 6, 2016, 2, pp. 9-11 (ISSN 2384-8987).

[3] Chomsky, Noam. 1957. Syntactic structures, The Hague/Paris, Mouton.

[4] Paul, Douglas B., and Janet Baker. "The design for the Wall Street Journal-based CSR corpus." Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992. 1992.

[5] Guo, Yufan, Anna Korhonen, Maria Liakata, Ilona Silins Karolinska, Lin Sun, and Ulla Stenius. 2010. 'Identifying the information structure of scientific abstracts: An investigation of three different schemes'. In Proceedings of the 2010 Workshop on Biomedical Natural Language Processing, pages 99–107, Association for Computational Linguistics, Stroudsburg,PA, USA.

[6] Marotta, Giovanna and Strik Lievers, Francesca (a cura di), Strutture linguistiche e dati empirici in diacronia e sincronia. Pisa, Pisa University Press ("Studi Linguistici Pisani" 8), 2017, 268 pp.

[7] Sprugnoli, Rachele and Sara Tonelli. 'Novel event detection and classification for historical texts.' Computational Linguistics 45, no. 2 (2019): 229-265.

[8] Sprugnoli, Rachele, Tommaso Caselli, Sara Tonelli, and Giovanni Moretti. 2017. 'The Content Types Dataset: a New Resource to Explore Semantic and Functional Characteristics of Texts.' In 15th Conference of the European Chapter of the Association for Computational Linguistics, Vol. 2, pp. 260-266. Association for Computational Linguistics.

[9] Sprugnoli, Rachele, 2018. 'Arretium or Arezzo? A Neural Approach to the Identification of Place Names in Historical Texts'. Negli Atti della Quinta Conferenza di Linguistica Computazionale Italiana, pp. 360-365. aAccademia University Press.

[4] https://en.wikisource.org/wiki/Main_Page

[5] http://www.gutenberg.org/

[6] https://github.com/tommasoc80/ContentTypes

[7] https://github.com/dhfbk/Histo

[8] https://github.com/dhfbk/Detection-of-place-names-in-historical-travel-writings

[9] https://stanfordnlp.github.io/CoreNLP/

[10] https://osf.io/g6w4e/?view_only=1d67431f4b0345f780d0e4c0b370d94a

[10] Teufel, Simone and Marc Moens. 2002. ‘Summarizing scientific articles: experiments with relevance and rhetorical status’. Computational linguistics, 28(4):409–445.
[11] Voghera, Miriam. 2001. ‘Teorie linguistiche e dati di parlato’. In Albano Leoni F., Stenta Krosbakken E., Sornicola R., Stromboli C. (a cura di), Dati empirici e teorie linguistiche, Atti del XXXIII Congresso Internazionale di Studi della Società di linguistica italiana (SLI), Bulzoni, Roma, pp. 75-95.
[12] Werlich, Egon, 1976. A text grammar of English. Quelle & Meyer.