WIT^3: IL CORPUS DEI SOTTOTITOLI MULTILINGUE DEGLI INTERVENTI ALLE CONFERENZE TED

Mauro Cettolo, Christian Girardi e Marcello Federico
FBK – Fondazione Bruno Kessler
Trento, Italy
{*cettolo,cgirardi,federico*}*@fbk.eu*

1. SOMMARIO

In questo lavoro viene presentato WIT^3, il sito web che abbiamo sviluppato per distribuire una versione pronta all'uso della collezione di sottotitoli multilingua degli interventi alle conferenze TED. Siamo persuasi che questa collezione rappresenti una risorsa preziosa per la comunità scientifica che si occupa di traduzione automatica, data la sua dimensione in continua crescita e data la sua varietà sia in termini di lingue sia di argomenti trattati. Infatti già ad oggi, giugno 2013, il sito TED raccoglie la registrazione di più di 2000 interventi che spaziano su tutto lo scibile umano, dalla tecnologia all'intrattenimento, dall'economia alla scienza; le trascrizioni in inglese sono già disponibili per la maggior parte delle registrazioni, mentre le traduzioni vengono via via aggiunte e al momento coprono fino a 100 lingue diverse.

La nostra ambizione è di fornire attraverso WIT^3 un servizio adeguato alla comunità scientifica distribuendo: (a) per un numero consistente di coppie di lingue il materiale per l'addestramento di sistemi statistici di traduzione e la loro valutazione, insieme a delle traduzioni generate automaticamente che possono fungere da riferimento; (b) i file originali del sito di TED con degli strumenti di elaborazione che consentono a chiunque di preparare autonomamente l'ambiente sperimentale per qualsiasi coppia di lingue.

2. INTRODUZIONE

I dati giocano un ruolo chiave nell'apprendimento automatico – noto in letteratura come Machine Learning – essendo essi la principale sorgente di informazione da cui inferire i valori dei parametri dei modelli matematici in uso.

Nella traduzione automatica statistica (statistical machine translation, SMT), l'apprendimento viene compiuto su testi paralleli, ovvero documenti, frasi o anche semplici frammenti di frasi accoppiati alle loro rispettive traduzioni in una o più lingue. È tipico che per addestrare adeguatamente i modelli di traduzione e di riordinamento di un sistema SMT sia necessario impiegare una grande quantità di dati paralleli, possibilmente nel dominio semantico di interesse.

Purtroppo, i dati paralleli sono una risorsa scarsa, disponibile solo per alcune coppie di lingue e per pochi domini, spesso molto specifici. Ad esempio, MultiUN (www.euromatrixplus.net/multi-un/) fornisce una quantità notevole di dati paralleli, ma per sole sei lingue; Europarl (www.statmt.org/europarl/) include la traduzione nella maggior parte delle lingue europee degli atti del Parlamento Europeo (fino a 50 milioni di parole); JRC-Acquis (langtech.jrc.ec.europa.eu/JRC-Acquis.html) comprende l'intero corpo della legislazione dell'Unione Europea che si applica agli stati membri, tradotta completamente o parzialmente in 22 lingue (da 30 a 60 milioni di parole per ciascuna lingua); altri corpora paralleli più piccoli per domini molto specifici si trovano in OPUS (opus.lingfil.uu.se) per alcune decine di lingue.

D'altro canto, è impensabile per i laboratori di ricerca coprire ogni possibile esigenza in termini di corpora paralleli ricorrendo a traduttori professionisti, dato il loro alto costo.

I dati disponibili sul sito di TED (www.ted.com) risultano quindi particolarmente preziosi per la comunità della traduzione automatica. TED è un'organizzazione nonprofit che invita "gli intellettuali ed i professionisti più brillanti a tenere il discorso della loro vita". Il sito rende disponibili, con licenza Creative Commons BY-NC-ND, le registrazioni audio-video dei migliori interventi con tanto di sottotitoli in inglese e la loro traduzione eseguita da volontari in diverse lingue. L'insieme dei sottotitoli rappresenta pertanto una risorsa parallela multilingue di valore inconfutabile, giacché cresce continuamente nel tempo (ad oggi, giugno 2013, il sito mette a disposizione le registrazioni di oltre 2000 interventi), include le traduzioni in decine e decine di lingue (vi sono interventi tradotti in 100 idiomi diversi), italiano incluso, e copre argomenti che spaziano su tutto lo scibile umano, rendendo la risorsa potenzialmente utile per qualsiasi applicazione.

Con l'obiettivo di rendere questo corpus di fruibilità immediata presso la comunità scientifica, abbiamo sviluppato WIT^3 (wit3.fbk.eu), un sito web che ospita una versione pronta all'uso di questa risorsa multilingue, dei benchmark di riferimento per la traduzione automatica e degli strumenti software per la gestione e manipolazione dei suoi testi.

Oltre che di per se, il sito di WIT^3 svolge un importante ruolo per IWSLT, il workshop internazionale per la traduzione del linguaggio parlato (International Workshop on Spoken Language Translation), che si tiene a cadenza annuale. A partire dall'edizione 2012, infatti, i dati per l'addestramento, la calibrazione e la valutazione di sistemi per la traduzione automatica dei discorsi TED, uno dei problemi proposti nella campagna di valutazione di IWSLT, vengono rilasciati attraverso il sito di WIT^3. Accanto alle risorse linguistiche e agli strumenti software, il sito rende disponibili anche le traduzioni automatiche generate da sistemi di base e la loro valutazione in termini di BLEU e TER, due delle metriche più comuni in uso nella traduzione automatica; in questo modo vengono forniti non solo ai partecipanti ma alla comunità intera dei risultati di riferimento con cui validare le prestazioni dei propri sistemi.

Il prossimo paragrafo fornisce la descrizione dettagliata del corpus dei discorsi di TED, del formato dei file e della procedura seguita per produrre l'allineamento a livello di frasi; vengono inoltre fornite delle statistiche sul corpus, con particolare riferimento alle lingue della campagna di valutazione IWSLT 2013, attualmente in corso; viene inoltre proposta un'analisi quantitativa della difficoltà di tradurre automaticamente i sottotitoli di TED. La relazione tra WIT^3 e IWSLT è oggetto del paragrafo 4, mentre quello successivo fornisce una panoramica delle caratteristiche salienti dei sistemi di base che abbiamo sviluppato quali riferimenti per le ultime campagne di valutazione di IWSLT, compreso quello per la traduzione tra l'inglese e l'italiano. L'articolo termina con la presentazione del sito web di WIT^3 (paragrafo 6).

3. IL CORPUS DEGLI INTERVENTI ALLE CONFERENZE TED

Gli interventi alle conferenze TED sono tenuti perlopiù in inglese e le registrazioni audio-video sottotitolate sono accessibili dal sito di TED. La gran parte dei sottotitoli sono stati tradotti da dei volontari in arabo, cinese, coreano, ebraico, francese, italiano, polacco, portoghese-brasiliano, rumeno e spagnolo. Per altre decine di lingue, le sottotitolazioni tradotte variano da un numero abbondantemente superiore alle mille (bulgaro, olandese, russo, tedesco, turco) a poche unità (una sola per 14 lingue tra le quali citiamo il latino e l'occitano). Va tenuto presente che i sottotitoli originali e le loro traduzioni sono segmentati sulla base

dell'audio e pertanto i singoli sottotitoli non corrispondono necessariamente a delle frasi di senso compiuto: può accadere sia che una frase sia spezzata su più sottotitoli sia che uno stesso sottotitolo contenga (frammenti di) frasi consecutive.

Per preparare i dati paralleli, dapprima si raccolgono i dati grezzi dal sito di TED, successivamente si raggruppano le trascrizioni e le traduzioni dello stesso discorso, poi si allineano i sottotitoli nelle diverse lingue ed infine si ricostruiscono le frasi di senso compiuto mantenendo l'allineamento. A ciascuno dei passi menzionati è dedicato uno dei prossimi paragrafi.

3.1. Raccolta dei dati grezzi dal sito di TED

I dati grezzi vengono raccolti dal sito di TED per mezzo di HLTWebManager (Girardi, 2011), un crawler sviluppato in Java da noi che permette di individuare e scaricare delle pagine web scritte in lingue differenti che verosimilmente sono traduzioni le une delle altre. Nei file HTML originali vengono individuati i sottotitoli ed alcuni dei metadati relativi al discorso, che vengono poi salvati in un file in formato XML definito dal DTD disponibile nel sito di WIT^3 (paragrafo 6).

Per ogni lingua viene generato un singolo file XML che include tutti i discorsi per i quali la sottotitolazione in quella lingua è disponibile. I discorsi all'interno di un XML sono delimitati dalle etichette `<file id=int>` e `</file>` e per ciascuno di essi vengono fornite, tra le altre, le seguenti informazioni:

`<url>`	l'indirizzo del documento HTML originale contenente il discorso
`<speaker>`	il nome del conferenziere
`<talkid>`	un identificativo numerico del discorso
`<transcript>`	i sottotitoli del discorso segmentati temporalmente
`<date>`	la data di svolgimento della conferenza
`<content>`	i sottotitoli del discorso

I campi `transcript` e `content` differiscono solo per la presenza nel primo dei riferimenti temporali usati per sincronizzare le visione dei sottotitoli durante la riproduzione audio-video.

Il `talkid` è un intero che identifica univocamente la trascrizione originale del discorso e tutte le sue traduzioni. È pertanto usato per accoppiare i sottotitoli in lingue diverse dello stesso discorso.

Altre etichette quali `description` (descrizione), `keywords` (parole chiave), `title` (titolo), il cui significato è perspicuo, forniscono meta-informazioni sul discorso utili ad esempio per raggruppare discorsi su argomenti simili, per l'archiviazione ed il successivo recupero di discorsi specifici, per la classificazione dei discorsi o per l'adattamento di modelli statistici.

3.2. Allineamento

Data una coppia di lingue, è facile selezionare i discorsi per i quali è disponibile la sottotitolazione in entrambe esse, sfruttando il campo `talkid` menzionato nel paragrafo 3.1.

Da ciascuno di questi discorsi, i sottotitoli nelle due lingue sono estratti dal campo `transcript` e accoppiati in ordine di apparizione. Viene quindi effettuata una serie di controlli euristici per scartare accoppiamenti presumibilmente errati: un discorso intero viene completamente scartato se vi è una discrepanza nel numero totale di sottotitoli per le due

lingue o nei riferimenti temporali dei singoli sottotitoli. Singoli sottotitoli accoppiati vengono rimossi se il rapporto tra le loro lunghezze è significativamente anomalo, assumendo una distribuzione normale del loro rapporto e un intervallo di confidenza del 95%.

Per farsi un'idea dell'impatto di questo filtraggio, nel caso della collezione inglese-francese esso comporta l'eliminazione di circa il 3% delle parole complessive.

Una volta che i sottotitoli sono allineati, vengono ricostruite delle frasi di senso compiuto concatenando (sincronicamente su entrambe le lingue) sottotitoli consecutivi fino a quando non si incontra un sottotitolo della seconda lingua che finisce con un segno di punteggiatura terminale. Ciò significa che i testi paralleli così creati possono includere: (a) righe che includono più frasi e (b) righe della prima lingua che non terminano con un segno di punteggiatura terminale.

	ar	de	en	es	fa	fr	it	nl	pl	pt-br	ro	ru	sl	tr	zh
ar	-	2,21	2,60	2,47	1,87	2,69	2,43	2,21	1,90	2,44	2,45	1,86	0,22	1,62	0,55
de	2,36	-	2,43	2,31	1,84	2,51	2,27	2,10	1,78	2,29	2,29	1,78	0,22	1,56	0,51
en	2,62	2,29	-	2,59	2,00	2,82	2,55	2,35	1,97	2,54	2,58	1,93	0,24	1,67	0,57
es	2,59	2,26	2,69	-	1,93	2,77	2,51	2,27	1,94	2,51	2,51	1,91	0,23	1,64	0,55
fa	1,31	1,21	1,37	1,26	-	1,37	1,25	1,25	0,97	1,26	1,25	1,05	0,19	0,88	0,29
fr	2,60	2,26	2,69	2,53	1,87	-	2,49	2,28	1,94	2,52	2,51	1,90	0,22	1,65	0,56
it	2,61	2,28	2,71	2,57	1,89	2,78	-	2,30	1,95	2,54	2,53	1,92	0,23	1,66	0,56
nl	2,37	2,09	2,49	2,32	1,88	2,53	2,29	-	1,77	2,31	2,32	1,79	0,23	1,54	0,51
pl	2,46	2,14	2,52	2,39	1,78	2,60	2,34	2,13	-	2,37	2,38	1,80	0,22	1,59	0,53
pt-br	2,57	2,25	2,65	2,51	1,88	2,75	2,47	2,26	1,92	-	2,50	1,87	0,23	1,64	0,55
ro	2,58	2,26	2,71	2,53	1,88	2,75	2,48	2,28	1,93	2,51	-	1,92	0,23	1,66	0,55
ru	2,21	1,98	2,27	2,15	1,77	2,34	2,11	1,99	1,64	2,12	2,15	-	0,22	1,46	0,48
sl	0,27	0,25	0,29	0,27	0,32	0,28	0,26	0,27	0,20	0,26	0,27	0,23	-	0,19	0,06
tr	2,27	2,04	2,32	2,20	1,74	2,41	2,17	2,02	1,72	2,19	2,20	1,72	0,21	-	0,49
zh	2,55	2,22	2,63	2,47	1,88	2,69	2,43	2,22	1,91	2,45	2,45	1,86	0,22	1,63	-

Tabella 1: Numero (milioni) di unità delimitate da spazi che occorrono complessivamente nei testi paralleli per alcune coppie di lingue. La matrice è simmetrica rispetto ai dati rappresentati, ovvero le caselle (`i,j`) e (`j,i`) riferiscono entrambe gli stessi dati paralleli della coppie di lingue `i`→`j` e `j`→`i`; i valori contenuti sono invece diversi in quanto vengono sempre forniti i valori relativi alla seconda lingua, ovvero quella della colonna. I nomi delle lingue sono rappresentati dai codici ISO 639-1.

3.3. Statistiche

Le statistiche aggiornate a giugno 2013 dicono che abbiamo raccolto un totale di circa 35 mila discorsi sottotitolati, originati dalla traduzione in più di 80 lingue di 1350 discorsi tenuti in inglese.

I testi che mettiamo a disposizione sono quelli originali, ovvero non applichiamo alcun processo di separazione della punteggiatura dalle parole, né di segmentazione in parole del testo nelle lingue che ne sono prive quali il cinese ed il giapponese. Pertanto, quando riportiamo le dimensioni dei corpora ci riferiamo al numero di unità comprese tra spazi, che quindi non necessariamente sono parole.

Per quanto riguarda il cinese, consideriamo solo i documenti scritti con caratteri in forma semplice, o caratteri semplificati, mentre scartiamo quelli scritti in caratteri tradizionali.

La distribuzione delle traduzioni tra le oltre 80 lingue è molto irregolare e conseguentemente risulta essere ancor più sparsa quelle tra tutte le oltre 3000 coppie possibili.

La tabella 1 mostra la dimensione dei testi paralleli che siamo in grado di generare per tutte le 210 possibili coppie delle 15 lingue coinvolte nella campagna di valutazione IWSLT 2013. Si tenga presente che gli organizzatori hanno proposto solo coppie che includono l'inglese come lingua sorgente o come lingua obiettivo; la tabella presenta invece anche i valori delle altre coppie a scopo di analisi.

Le coppie con l'inglese includono sempre sulla sua parte almeno 2 milioni di unità, con l'eccezione del persiano (farsi) e dello sloveno. Queste due lingue sono state proposte quali rappresentanti di gruppi della famiglia linguistica indoeuropea diversi da quello dell'inglese (germanico questo, iranico e slavo gli altri due) con al contempo una quantità di dati sensibilmente inferiore agli altri casi, proprio per quantificare l'impatto in queste condizioni della scarsa disponibilità di materiale per l'addestramento statistico.

Per quanto riguarda le coppie senza l'inglese, la situazione sulla quantità di dati paralleli ricalca quella della campagna di valutazione, ovvero in generale siamo riusciti a generare circa 2 milioni di unità, in alcuni casi anche più di 2 milioni e mezzo (ad esempio col francese, l'italiano ed il rumeno), con l'eccezione delle coppie che includono il persiano e lo sloveno: per esse ad ogni modo la quantità di dati paralleli generati è sempre all'incirca equivalente a quella disponibile con l'inglese, pertanto sotto questo punto di vista la situazione non peggiora.

In generale, quindi, se la coppia di lingue a cui siamo interessati comprende almeno una lingua per la quale molti discorsi sono stati trascritti, verosimilmente ci si può aspettare che la quantità di dati paralleli generabili sia prossima a quella della lingua meno rappresentata in TED. Se invece entrambe le lingue sono poco rappresentate, allora è probabile che siano pochi i discorsi trascritti in entrambe esse e quindi altrettanto pochi i dati paralleli generabili.

3.4. Approfondimenti

Quanto è difficile tradurre i discorsi TED? È chiaro che una risposta assoluta non esiste, ma si può tentare di fornirne una relativa. Prendiamo come riferimento i risultati ottenuti dai partecipanti alla campagna di valutazione IWSLT 2012 (Federico, Cettolo, Bentivogli, Paul, & Stüker, 2012). I migliori punteggi per la traduzione inglese→francese misurati con metriche automatiche rispetto ad una singola frase di riferimento, riportati in tabella 8, sono paragonabili a quelli dei migliori sistemi della campagna di valutazione WMT 2011 (Callison-Burch, Koehn, Monz, & Zaidan, 2011) nel dominio delle agenzie di stampa. Questo confronto è particolarmente significativo data l'analogia delle condizioni sperimentali: quantità equivalente di dati di addestramento nel dominio di interesse e stessi dati fuori da quel dominio. D'altra parte, i punteggi sulla traduzione dei discorsi TED dall'arabo e dal cinese verso l'inglese sono nettamente peggiori di quelli ottenuti sui notiziari dai migliori sistemi nell'ultima campagna di valutazione NIST[1] per le stesse coppie di lingue; in questo caso però non solo le condizioni di addestramento sono molto diverse in quanto NIST fornisce una molto più consistente mole di dati, ma anche il calcolo dei punteggi viene effettuato rispetto a dei riferimenti multipli, non singoli.

[1] `www.itl.nist.gov/iad/mig/tests/mt/2009/ResultsRelease/progress.html`

Ad ogni modo la prima considerazione che ci sentiamo di fare è che tradurre i discorsi TED è difficile quanto tradurre i notiziari o le agenzie di stampa a partità di lingue; nel caso di coppie di lingue particolarmente distanti, il degrado di prestazioni è sensibile rispetto a coppie più vicine, ma potrebbe essere mitigato aumentando la quantità di dati di addestramento.

Aldilà delle prestazioni dei sistemi di traduzione automatici, la difficoltà di tradurre automaticamente una lingua in un'altra può essere predetta dalla perplessità (PP) e dal tasso di parole fuori dal dizionario (out-of-vocabulary word rate, OOV) del modello del linguaggio (language model, LM) a *n*-grammi della lingua obiettivo. Quando questi valori sono calcolati su testi del dominio, essi forniscono un'indicazione di quanto il compito di traduzione è intrinsecamente difficile; se calcolati su testi fuori dal dominio, essi forniscono un'indicazione sull'utilità in addestramento di questi testi nonostante la loro distanza dal dominio di interesse.

Analizziamo come caso-studio la traduzione inglese→francese del corpus di valutazione `tst2011` (paragrafo 5.1.1). Abbiamo stimato dei LM a 5-grammi su testi francesi messi a disposizione per l'addestramento, in particolare:

- TED: testi monolingua francesi di TED; è l'unico testo nel dominio di interesse
- NC: parte francese del corpus parallelo inglese-francese News Commentary
- EPPS: parte francese del corpus parallelo inglese-francese Europarl
- MultiUN: parte francese del corpus parallelo multiplo MultiUN.

Sono quindi stati calcolati i valori, riportati in tabella 2, di PP/OOV della parte francese di `tst2011` su ciascun LM; la tabella specifica anche il numero di parole usate per l'addestramento dei modelli.

corpus	dimensione del corpus	PP	%OOV
TED	2,35M	103,8	1,67
NC	3,36M	266,8	2,83
EPPS	56,2M	200,3	1,79
MultiUN	402,8M	288,2	1,21
unione	464,7M	150,8	0,72

Tabella 2: PP e %OOV del testo di valutazione `tst2011` rispetto a quattro LM a 5-grammi stimati su corpora nel dominio e fuori dal dominio. L'ultima riga riporta i valori calcolati sul LM costruito sull'unione dei quattro corpora. Il simbolo `M` sta per 10^6.

Possiamo dedurre quanto segue:

- il corpus nel dominio ottiene la miglior PP, nonostante sia il più piccolo; questo dimostra che sebbene i discorsi TED trattino temi piuttosto diversi, l'ambito comune induce i conferenzieri ad usare un linguaggio simile
- i discorsi TED sono linguisticamente piuttosto distanti dagli altri generi considerati in questa analisi, ovvero i notiziari, gli atti del parlamento europeo e le risoluzioni dell'assemblea generale delle Nazioni Unite. È piuttosto inaspettato che EPPS risulti essere più vicino ai discorsi TED dei notiziari, ma la differenza potrebbe essere dovuta alla dimensione dei due corpora piuttosto che alla loro natura

- l'OOV rispetto ai corpora fuori dal dominio sembra essere principalmente correlata alla loro dimensione; è interessante osservare che la OOV può essere più che dimezzata se tutti i dati fuori dal dominio vengono aggiunti a quelli nel dominio (si faccia riferimento alla riga `unione`), il che dimostra come lo sfruttamento di ogni dato disponibile possa comunque essere benefico.

I valori appena commentati fanno riferimento al testo di valutazione considerato nel suo insieme, ma esso si compone di più discorsi e ci si potrebbe chiedere se vi sia una differenza tra un discorso e l'altro. La tabella 3 fornisce dei valori calcolati a livello del singolo discorso sia dei testi di valutazione sia di quelli di addestramento; nello specifico, vengono riportati i valori della media μ di PP e OOV, la loro deviazione standard σ, ed i valori minimi e massimi. Per quel che riguarda il testo di valutazione, i valori sono calcolati sul LM stimato sui dati di addestramento; i valori riportati per i dati di addestramento sono stati invece calcolati usando uno schema di validazione incrociato (1-fold cross validation).

		μ	σ	[min;max]
tst2011	PP	103,7	19,7	[68,9;132,0]
	%OOV	1,55	0,46	[0,91;2,37]
addestramento	PP	130,2	49,3	[38,8;505,7]
	%OOV	1,76	1,04	[0,00;15,79]

Tabella 3: Media, deviazione standard e valori minimo e massimo di PP e %OOV dei discorsi TED inclusi nei dati di valutazione e in quelli di addestramento.

Risulta quanto segue:

- in media, i valori di PP e OOV dei discorsi selezionati per la valutazione sono inferiori a quelli dei discorsi utilizzabili per l'addestramento; questo è verosimilmente dovuto alla cura posta nella selezione, controllo e pulizia dei discorsi per la valutazione, che ne ha significativamente migliorato la qualità rispetto a quella che caratterizza i dati generati in maniera completamente automatica per l'addestramento
- gli intervalli [min,max] dei valori osservati di PP e OOV sono piuttosto ampi; questo significa che i discorsi possono linguisticamente differire fra di loro in maniera significativa e di conseguenza anche le prestazioni di traduzione automatica su di essi.

4. WIT^3 PER LE CAMPAGNE DI VALUTAZIONE DI IWSLT

Il convegno internazionale per la traduzione del parlato (International Workshop on Spoken Language Translation, IWSLT) è un evento annuale con associata una campagna di valutazione di traduzione automatica della lingua parlata. IWSLT propone ogni anno stimolanti problemi di ricerca ed una infrastruttura sperimentale aperta rivolta alla comunità scientifica che lavora alla traduzione automatica del parlato e del testo scritto.

Quando parliamo di "infrastruttura sperimentale aperta" ci riferiamo ad un aspetto caratterizzante IWSLT che si declina in due modi: da una parte le risorse linguistiche preparate specificatamente per la campagna di valutazione vengono messe a disposizione dei partecipanti gratuitamente; dall'altra tutti i risultati e gli articoli presentati al convegno vengono pubblicati sul sito dello stesso.

Nell'edizione 2010 della campagna di valutazione (Paul, Federico, & Stücker, 2010), alla consueta traduzione di testi relativi al dominio dei viaggi del corpus BTEC (Takezawa, Kikui, Mizushima, & Sumita, 2007), fu affiancata per la prima volta la traduzione dei discorsi TED. Questo è rimasto l'unico esercizio di traduzione delle edizioni 2011 (Federico, Bentivogli, Paul, & Stüker, 2011) e 2013 attualmente in corso,[2] e quello principale dell'edizione 2012 dove fu accoppiato alla traduzione dal cinese all'inglese di testi relativi ai giochi olimpici di Pechino (Federico et al., 2012).

Come abbiamo visto nel paragrafo 3, i dati nel dominio di TED distribuiti per la stima dei modelli statistici contano approssimativamente 2 milioni di unità, quantità che varia a seconda della coppia di lingue e che tendenzialmente cresce di anno in anno con l'aumentare dei discorsi trascritti e tradotti rilasciati da TED. Accanto ai dati nel dominio, vengono distribuiti anche dati fuori dal dominio sia paralleli sia monolingua, quali i testi delle risoluzioni delle Nazioni Unite, gli atti del parlamento europeo e agenzie di stampa e notiziari.

A partire dall'edizione 2012 di IWSLT, i dati nel dominio di TED vengono distribuiti attraverso il sito di WIT^3 con delle edizioni speciali del corpus preparate appositamente. In aggiunta alle edizioni speciali, il sito rilascia anche edizioni di WIT^3 che coprono altre coppie di lingue non proposte in IWSLT.

5. SISTEMI DI RIFERIMENTO

In questo paragrafo presentiamo alcuni sistemi di base per la traduzione automatica dei discorsi TED e le loro prestazioni. Lo scopo è quello di fornire ai ricercatori interessati dei riferimenti con cui possano validare i loro sistemi ed i loro progressi.

Oltre alle coppie di lingue proposte nella campagna di valutazione IWSLT 2012, forniamo anche le prestazioni per il sistema italiano→inglese costruito sui dati rilasciati per la campagna 2013 attualmente in corso, immaginando di fare cosa gradita al lettore.

5.1. Le coppie di lingue proposte a IWSLT 2012

5.1.1. Dati

Gli esperimenti sono stati effettuati sui dati forniti dagli organizzatori della campagna di valutazione di IWSLT 2012 per la traduzione dei discorsi TED;[3] le direzioni di traduzione proposte sono:

`ufficiali`: inglese→francese e arabo→inglese

`opzionali`: {tedesco, olandese, polacco, portoghese-brasiliano, rumeno, russo, slovacco, sloveno, turco, cinese} → inglese

Per lo sviluppo dei sistemi di traduzione abbiamo utilizzato unicamente i dati nel dominio TED, ovvero nessun corpora supplementare fuori dal dominio comunque messo a disposizioni dagli organizzatori è stato impiegato.

I testi sono stati pre-elaborati in maniera diversa a seconda della lingua: l'arabo ed il cinese sono stati segmentati per mezzo di AMIRA (Diab, Hacioglu, & Jurafsky, 2004) e del segmentatore per il cinese sviluppato presso l'università di Stanford (Tseng, Chang, Andrew, Jurafsky, & Manning, 2005), rispettivamente; la separazione della punteggiatura dalle parole

[2] www.iwslt2013.org

[3] hltc.cs.ust.hk/iwslt/index.php/evaluation-campaign/ted-task.html

lingua	frasi	unità	dizionario
en	142k	2,82M	54,8k
fr	143k	3,01M	67,3k

Tabella 4: Risorse monolingua per le coppie ufficiali di lingue: per ciascuna lingua sono specificati il numero di frasi di cui si compone il corpus, il numero totale di unità e di unità distinte (dizionario) che vi occorrono. I simboli `M` e `k` stanno per 10^6 e 10^3 rispettivamente.

coppia	corpus	lingua	frasi	unità	dizionario	discorsi
en→fr	addestramento	en	141k	2,77M	54,3k	1.029
		fr		2,91M	66,9k	
	dev2010	en	934	20,1k	3,4k	8
		fr		20,3k	3,9k	
	tst2010	en	1.664	32,0k	3,9k	11
		fr		33,8k	4,8k	
	tst2011	en	818	14,5k	2,5k	8
		fr		15,6k	3,0k	
	tst2012	en	1.124	21,5k	3,1k	11
		fr		23,5k	3,7k	
ar→en	addestramento	ar	138k	2,54M	89,7k	1.015
		en		2,73M	53,9k	
	dev2010	ar	934	18,3k	4,6k	8
		en		20,1k	3,4k	
	tst2010	ar	1.664	29,3k	6,0k	11
		en		32,0k	3,9k	
	tst2011	ar	1.450	25,6k	5,6k	16
		en		27,0k	3,7k	
	tst2012	ar	1.704	27,8k	6,1k	15
		en		30,8k	4,1k	

Tabella 5: Risorse bilingue per le coppie ufficiali di lingue: in aggiunta ai valori forniti per le risorse monolingue (si legga la didascalia alla tabella 4), viene fornito anche il numero di discorsi contenuti nei vari corpora.

nei testi in altre lingue è stata invece effettuata dallo strumento rilasciato all'uopo con il corpus Europarl (Koehn, 2005).

Per l'addestramento dei modelli, le risorse linguistiche nel dominio messe a disposizione comprendono sia testi paralleli sia monolingua. Per la calibrazione dei sistemi delle due coppie di lingue ufficiali, sono stati rilasciati gli stessi testi delle edizioni precedenti, chiamati `dev2010` e `tst2010`. Per la valutazione, un nuovo insieme di discorsi è stato selezionato per definire il `tst2012` che è stato affiancato al `tst2011`, usato a sua volta per la valutazione dei progressi effettuati dalla comunità scientifica a distanza di un anno.

Per le coppie opzionali sono stati preparati analoghi insiemi di calibrazione e valutazione, sulla base degli stessi discorsi inclusi negli insiemi definiti per la coppia arabo→inglese.

Alcune statistiche sui dati rilasciati relativi alle coppie ufficiali sono riportati nelle tabelle 4 e 5; i valori si riferiscono ai testi pre-elaborati con gli strumenti summenzionati.

5.1.2. Prestazioni

I sistemi di traduzione di riferimento sono statistici e sviluppati con il software open-source Moses (Koehn et al., 2007). I modelli di traduzione e riordinamento sono stati addestrati sui dati bilingue; considerando la relativa esiguità della quantità di dati monolingua a disposizione, ci siamo limitati ad addestrare col pacchetto IRSTLM (Federico, Bertoldi, & Cettolo, 2008) dei modelli del linguaggio a 4-grammi, alle cui distribuzioni abbiamo applicato la tecnica Kneser-Ney (Chen & Goodman, 1999). La calibrazione dei sistemi, ovvero la stima dei pesi dell'interpolazione log-lineare dei vari modelli, è stata effettuata sul `dev2010` con la tecnica iterativa del MERT disponibile nel pacchetto Moses. Le prestazioni vengono fornite in termini di metriche standard quali BLEU, METEOR e TER, calcolati con `MultEval` (Clark, Dyer, Lavie, & Smith, 2011).

	%BLEU	σ	MTR	σ	TER	σ
en→fr						
dev2010	26,28	0,59	47,57	0,47	56,80	0,70
tst2010	28,74	0,47	49,63	0,37	51,30	0,47
tst2011	34,95	0,70	54,53	0,51	44,11	0,60
tst2012	34,89	0,61	54,68	0,44	43,35	0,50
ar→en						
dev2010	24,70	0,54	48,66	0,39	55,41	0,59
tst2010	23,64	0,45	47,61	0,34	57,16	0,50
tst2011	22,66	0,49	46,37	0,37	60,27	0,59
tst2012	24,05	0,44	48,62	0,31	54,72	0,43

Tabella 6: Prestazioni in termini di %BLEU, METEOR (MTR) e TER e loro deviazione standard (σ) dei sistemi di riferimento sulle coppie ufficiali di lingue.

Le tabelle 6 e 7 mostrano i valori di %BLEU, METEOR e TER dei sistemi di riferimento sviluppati per le coppie di lingue ufficiali e opzionali; sono stati calcolati tenendo conto degli eventuali errori sia sulla punteggiatura sia sulle maiuscole.

Oltre ai punteggi misurati sul `dev2010` dopo l'ultima iterazione del MERT, vengono riportati quelli relativi al secondo corpus di calibrazione `tst2010` e ai due corpora di valutazione `tst2011` e `tst2012`.

Sebbene i nostri modelli siano stati addestrati sui soli dati nel dominio e con una configurazione standard di Moses, i risultati in termini di BLEU e TER escono piuttosto bene dal confronto con quelli ottenuti dai partecipanti su `tst2011` e `tst2012` (Federico et al., 2012), i cui intervalli, limitatamente alle coppie ufficiali e al cinese→inglese, sono riportati in tabella 8. I valori di METEOR non sono confrontabili con quelli forniti ufficialmente dagli organizzatori a causa di diverse impostazioni di calcolo.

5.2. Il sistema di traduzione italiano-inglese

La campagna di valutazione 2013 include l'italiano-inglese, in entrambe le direzione, come coppia di lingue opzionale. Abbiamo quindi prelevato i dati distribuiti dagli organizzatori, costruito i due sistemi di traduzione di riferimento secondo lo schema definito nei paragrafi

	%BLEU	σ	MTR	σ	TER	σ
de→en						
dev2010	28,14	0,60	52,83	0,40	50,37	0,57
tst2010	26,18	0,48	50,86	0,34	52,59	0,50
tst2011	30,28	0,51	55,00	0,32	47,86	0,47
tst2012	26,55	0,48	50,99	0,32	52,42	0,46
nl→en						
dev2010	23,79	0,62	47,04	0,49	57,14	0,64
tst2010	31,23	0,48	54,62	0,32	47,90	0,45
tst2011	33,45	0,55	56,31	0,36	45,11	0,49
tst2012	29,89	0,46	53,16	0,31	47,60	0,42
pl→en						
dev2010	20,56	0,58	44,74	0,46	62,47	0,67
tst2010	15,27	0,36	40,03	0,31	69,95	0,47
tst2011	18,68	0,42	43,64	0,32	65,42	0,53
tst2012	15,89	0,39	39,11	0,32	68,56	0,48
pt-br→en						
dev2010	33,57	0,64	56,06	0,41	45,53	0,57
tst2010	35,27	0,47	58,85	0,31	43,01	0,43
tst2011	38,56	0,54	61,26	0,32	39,87	0,45
tst2012	40,74	0,50	62,09	0,29	37,96	0,40
ro→en						
dev2010	29,30	0,57	53,26	0,40	49,54	0,56
tst2010	28,18	0,47	52,32	0,33	51,13	0,46
tst2011	32,46	0,52	55,92	0,34	45,99	0,48
tst2012	29,08	0,48	52,73	0,33	50,32	0,45
ru→en						
dev2010	17,37	0,50	41,63	0,40	66,96	0,60
tst2010	16,82	0,37	41,93	0,29	66,28	0,47
tst2011	19,11	0,42	43,82	0,32	62,63	0,49
tst2012	17,44	0,39	41,73	0,31	63,94	0,43
sk→en						
dev2012	19,23	0,42	42,65	0,32	62,03	0,46
tst2012	21,79	0,58	45,01	0,41	58,28	0,55
sl→en						
dev2012	15,90	0,45	40,16	0,36	67,23	0,53
tst2012	14,33	0,39	39,42	0,33	69,20	0,50
tr→en						
dev2010	11,13	0,40	36,29	0,37	78,25	0,54
tst2010	12,13	0,32	37,87	0,27	75,56	0,45
tst2011	13,23	0,37	39,21	0,30	74,00	0,49
tst2012	12,45	0,33	38,76	0,29	73,63	0,43
zh→en						
dev2010	9,62	0,39	33,97	0,36	82,47	1,01
tst2010	11,39	0,32	36,80	0,28	75,99	0,76
tst2011	14,13	0,39	39,62	0,32	65,02	0,42
tst2012	12,33	0,33	37,67	0,30	67,80	0,39

Tabella 7: Prestazioni in termini di %BLEU, METEOR (MTR) e TER e loro deviazione standard (σ) dei sistemi di riferimento sulle coppie opzionali di lingue.

precedenti ed effettuato le misurazioni sulla qualità delle traduzioni automatiche degli insiemi di valutazione. Le statistiche sui dati ed i risultati sono riportati nelle tabelle 9 e 10, rispettivamente.

Confrontando i punteggi ottenuti nelle due direzioni chiaramente emerge la maggior difficoltà a tradurre dall'inglese all'italiano rispetto alla direzione opposta, fatto atteso visto che è ciò che tipicamente viene osservato quando si traduce tra lingue caratterizzate da diversi gradi di flessione.

Dalla comparazione coi risultati ottenuti sulle altre coppie di lingue e riportati nelle tabel-

coppia	corpus	%BLEU	MTR	TER
en→fr	tst2011	31,43–39,00	62,92–67,73	49,67–40,88
	tst2012	32,93–40,65	64,34–69,21	47,77–38,82
ar→en	tst2011	18,00–27,29	58,18–62,11	59,20–54,08
	tst2012	19,32–29,32	61,14–65,71	54,01–48,18
zh→en	tst2011	13,74–17,20	48,01–52,21	65,77–62,86
	tst2012	12,04–15,08	45,62–49,76	67,82–65,05

Tabella 8: Intervallo dei punteggi ufficiali ottenuti dai partecipanti alla campagna di valutazione IWSLT 2012 sugli insiemi `tst2011` e `tst2012`.

le 6 e 7 risulta poi che sull'italiano si ottengono punteggi tutto sommato alti, il che significa che la traduzione da e verso l'inglese della nostra lingua è sicuramente più agevole di quanto possa risultare quella di altre lingue quali le slave, il turco e soprattutto il cinese.

Riportiamo qui di seguito le traduzioni automatiche in entrambe le direzioni (`EN/IT_HYP`) di tre frasi pronunciate durante uno dei discorsi inclusi nel `tst2010`. Le frasi di riferimento corrette (`EN/IT_REF`) sono rispettivamente la trascrizione manuale di ciò che è stato effettivamente pronunciato in inglese e la sua traduzione manuale in italiano. Le `IT_HYP` sono state ottenute traducendo automaticamente le `EN_REF` col sistema en→it; viceversa, le `EN_HYP` sono state ottenute traducendo automaticamente le `IT_REF` col sistema it→en.

`IT_REF`: Qui vedete delle tipiche statistiche di mortalità organizzate per età .

`IT_HYP`: Quello che vediamo qui è un tipico grafico mortalità organizzato per età .

`EN_HYP`: Here you see of the typical statistics of organized mortality by age .

`EN_REF`: So what we 're looking at right here is a typical mortality chart organized by age .

`IT_REF`: Questo strumento che uso è un piccolo esperimento .

`IT_HYP`: Questo strumento che sto usando questo piccolo esperimento .

`EN_HYP`: This is a tool that we use is a little experiment .

`EN_REF`: This tool that I 'm using here is a little experiment .

`IT_REF`: Si chiama Pivot , e con Pivot posso scegliere di filtrare una sola causa di morte , diciamo gli incidenti .

`IT_HYP`: Si chiama Pivot Pivot , e con quello che possiamo fare è posso scegliere di filtrare in una particolare causa di morti , diciamo , un incidente .

`EN_HYP`: It 's called Pivot , and with Pivot I can choose to filter a single cause of death , let 's say accidents .

`EN_REF`: It 's called Pivot , and with Pivot what I can do is I can choose to filter in one particular cause of deaths – say , accidents .

corpus		lingua	frasi	unità	dizionario	discorsi
addestramento	monolingua	en	159k	3,19M	58,3k	1.210
	monolingua	it	159k	2,92M	84,1k	1.209
	bilingue	en	154k	3,06M	83,0k	1.163
		it		2,81M	66,9k	
dev2010		en	887	20,1k	3,4k	8
		it		17,6k	4,1k	
tst2010		en	1.529	31,0k	3,8k	10
		it		28,2k	5,0k	
tst2011		en	1.435	27,1k	3,7k	16
		it		24,3k	4,8k	
tst2012		en	1.704	30,8k	4,1k	15
		it		27,8k	5,4k	

Tabella 9: Risorse linguistiche per la coppia italiano-inglese rilasciate per la campagna di valutazione IWSLT 2013.

	%BLEU	σ	MTR	σ	TER	σ
en→it						
dev2010	22,72	0,57	39,48	0,54	56,95	0,60
tst2010	22,30	0,43	39,40	0,42	56,03	0,46
tst2011	23,29	0,51	40,06	0,49	56,67	0,54
tst2012	23,97	0,45	40,44	0,44	55,83	0,47
it→en						
dev2010	28,65	0,62	45,77	0,56	51,18	0,60
tst2010	28,14	0,44	45,99	0,41	51,93	0,46
tst2011	28,60	0,55	45,76	0,49	51,80	0,51
tst2012	28,08	0,46	45,58	0,42	52,04	0,44

Tabella 10: Prestazioni in termini di %BLEU, METEOR (MTR) e TER e loro deviazione standard (σ) dei sistemi di riferimento per la coppia italian-inglese nelle due direzioni.

Nonostante gli evidenti errori che mostrano quanto la traduzione automatica sia ancora perfettibile, il senso delle frasi viene mantenuto piuttosto fedelmente permettendone una facile comprensione.

6. IL SITO WEB DI WIT^3

L'indirizzo web di WIT^3 è:

`http://wit3.fbk.eu`

Il sito ospita diverse edizioni del corpus, comprese le speciali per le ultime due campagne di valutazione IWSLT, quella del 2012 e quella attualmente in corso del 2013. Le edizioni ordinarie hanno i seguenti link attivi:

`Plain texts for MT`: pagina caratterizzata da una tabella che ricorda grossomodo la tabella 1; ciascuna casella punta ad un archivio che contiene i dati paralleli e monolingua per l'addestramento e, se specificato attraverso una colorazione particolare della casella, anche dei corpora appositamente creati per la calibrazione e la valutazione dei sistemi di traduzione automatica

`Talks in XML format`: l'insieme dei file XML aggiornati (paragrafo 3.1), ognuno contenente tutti i discorsi sottotitolati in una certa lingua

`MT baseline results`: una tabella che fornisce sia le prestazioni di traduzione automatica ottenute con dei sistemi di riferimento sui testi di valutazione specificati sopra, sia le traduzioni medesime

`Tools`: l'insieme di strumenti software per la manipolazione dei dati TED; maggiori dettagli sono riportati di seguito.

Gli strumenti software rilasciati con ciascuna edizione del corpus sono degli script `perl`:

`find-common-talks.pl`: dati i file XML di due lingue, esso determina l'insieme dei `talkid` (si vedano i paragrafi 3.1 e 3.2) per i quali è disponibile la sottotitolazione in entrambe quelle lingue

`filter-talks.pl`: dato un file XML, esso seleziona i sottotitoli dei discorsi i cui `talkid` sono passati come parametro

`ted-extract-par.pl`: data una coppia di file XML, estrae i testi del campo `transcript` (paragrafi 3.1, 3.2) dei discorsi in comune, allineati a livello di singolo sottotitolo

`ted-extract-mono.pl`: dato un file XML, estrae il testo dal campo `transcript` (paragrafi 3.1, 3.2)

`rebuild-sent.pl`: ricostruisce le frasi a partire dai sottotitoli singoli (paragrafo 3.2).

Sfruttando i file XML e questi strumenti software, è possibile estrarre i dati paralleli e monolingua per qualsiasi coppia di lingue. Stabilendo un opportuno partizionamento dei `talkid` comuni ad una coppia di lingue, è poi possibile generare degli insiemi disgiunti di discorsi da utilizzare per l'addestramento dei modelli, per la calibrazione del sistema e per la sua valutazione.

Facciamo infine notare che il campo `url` (paragrafo 3.1) permette di indirizzare direttamente il file HTML originale di ciascun discorso, dando modo a chiunque di costruirsi la propria risorsa linguistica basata sui discorsi TED.

RINGRAZIAMENTI

Questo lavoro è finanziato da EU-Bridge (FP7-ICT-2011-7), un progetto del 7^o programma quadro della Commissione Europea.

BIBLIOGRAFIA

Callison-Burch, C., Koehn, P., Monz, C., & Zaidan, O. (2011), Findings of the 2011 Workshop on Statistical Machine Translation, in Proceedings of the Workshop on Statistical Machine Translation, Edinburgh, Scotland, 22–64.

Chen, S. F., & Goodman, J. (1999), An empirical study of smoothing techniques for language modeling, Computer Speech and Language, Vol. 4, no. 13, 359–393.

Clark, J., Dyer, C., Lavie, A., & Smith, N. (2011), Better hypothesis testing for statistical machine translation: controlling for optimizer instability, in Proceedings of the Association for Computational Lingustics, Portland, Oregon, USA,176-181.

Diab, M., Hacioglu, K., & Jurafsky, D. (2004), Automatic tagging of Arabic text: from raw text to base phrase chunks, in HLT-NAACL 2004: Short Papers, Boston, Massachusetts, USA, 149–152.

Federico, M., Bentivogli, L., Paul, M., & Stüker, S. (2011), Overview of the IWSLT 2011 evaluation campaign, in Proceedings of the International Workshop on Spoken Language Translation, San Francisco, California, USA, 11–27.

Federico, M., Bertoldi, N., & Cettolo, M. (2008), IRSTLM: an Open Source Toolkit for Handling Large Scale Language Models, in Proceedings of Interspeech, Brisbane, Australia, 1618–1621.

Federico, M., Cettolo, M., Bentivogli, L., Paul, M., & Stüker, S. (2012), Overview of the IWSLT 2012 evaluation campaign, in Proceedings of the International Workshop on Spoken Language Translation, Hong Kong, 12–33.

Girardi, C. (2011). The HLT Web Manager. (Report No. 23969). Trento, Italy: Fondazione Bruno Kessler (FBK). (https://wit3.fbk.eu/tools/WebManager_Manual.pdf)

Koehn, P. (2005), Europarl: a parallel corpus for statistical machine translation, in Proceedings of the Machine Translation Summit (MT Summit), Phuket, Thailand, 79–86.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., ... Herbst, E. (2007), Moses: open source toolkit for statistical machine translation, in Proceedings of the Association for Computational Linguistics Companion Volume of the Demo and Poster Sessions, Prague, Czech Republic, 177–180.

Paul, M., Federico, M., & Stücker, S. (2010), Overview of the IWSLT 2010 evaluation campaign, in Proceedings of the International Workshop on Spoken Language Translation, Paris, France, 3–27.

Takezawa, T., Kikui, G., Mizushima, M., & Sumita, E. (2007), Multilingual spoken language corpus development for communication research, Computational Linguistics and Chinese Language Processing, Vol. 12, no. 3, 303-324.

Tseng, H., Chang, P., Andrew, G., Jurafsky, D., & Manning, C. (2005), A conditional random field word segmenter, in Proceedings of the SIGHAN Workshop on Chinese Language Processing, Jeju Island, Korea.