Adattamento al Progetto dei Modelli di Traduzione Automatica nella Traduzione Assistita

Mauro Cettolo **Nicola Bertoldi** **Marcello Federico**
FBK - Fondazione Bruno Kessler
Trento, Italy
cognome@fbk.eu

Abstract

Italiano. L'integrazione della traduzione automatica nei sistemi di traduzione assistita è una sfida sia per la ricerca accademica sia per quella industriale. Infatti, i traduttori professionisti percepiscono come cruciale l'abilità dei sistemi automatici di adattarsi al loro stile e alle loro correzioni. In questo articolo proponiamo uno schema di adattamento dei sistemi di traduzione automatica ad uno specifico documento sulla base di una limitata quantità di testo, corretto manualmente, pari a quella prodotta giornalmente da un singolo traduttore.

English. *The effective integration of MT technology into computer-assisted translation tools is a challenging topic both for academic research and the translation industry. Particularly, professional translators feel crucial the ability of MT systems to adapt to their feedback. In this paper, we propose an adaptation scheme to tune a statistical MT system to a translation project using small amounts of post-edited texts, like those generated by a single user in even just one day of work.*

1 Introduzione

Nonostante i significativi e continui progressi, la traduzione automatica (TA) non è ancora in grado di generare testi adatti alla pubblicazione senza l'intervento umano. D'altra parte, molti studi hanno confermato che nell'ambito della traduzione assistita la correzione di testi tradotti automaticamente permette un incremento della produttività dei traduttori professionisti (si veda il paragrafo 2). Questa applicazione della TA è tanto più efficace quanto maggiore è l'integrazione del sistema di traduzione automatico nell'intero processo di traduzione, che può essere ottenuta specializzando il sistema sia al particolare testo da tradurre sia alle caratteristiche dello specifico traduttore e alle sue correzioni. Nell'industria della traduzione, lo scenario tipico è quello di uno o più traduttori che lavorano per alcuni giorni su un dato *progetto di traduzione*, ovvero su un insieme di documenti omogenei. Dopo un giorno di lavoro, le informazioni contenute nei testi appena tradotti e le correzioni apportate dai traduttori possono essere immesse nel sistema automatico con l'obiettivo di migliorare la qualità delle traduzioni automatiche proposte il giorno successivo. Chiameremo questo processo *adattamento al progetto*. L'adattamento al progetto può essere ripetuto quotidianamente fino al termine del lavoro, in modo da sfruttare al meglio tutte le informazioni che implicitamente i traduttori mettono a disposizione del sistema.

Questo articolo presenta uno dei risultati del progetto europeo MateCat,[1] nel cui ambito abbiamo sviluppato un sistema per la traduzione assistita basato sul Web integrante un modulo di TA che si auto-adatta allo specifico progetto. Gli esperimenti di validazione che andremo ad illustrare sono stati effettuati su quattro coppie di lingue, dall'inglese all'italiano (IT), al francese (FR), allo spagnolo (ES) e al tedesco (DE), e in due domini, tecnologie dell'informazione e della comunicazione (TIC) e legale (LGL).

Idealmente, i metodi di adattamento proposti dovrebbero essere valutati misurando il guadagno in termini di produttività su progetti di traduzione reali. Pertanto, per quanto possibile, abbiamo eseguito delle *valutazioni sul campo* in cui dei traduttori professionisti hanno corretto le traduzioni ipotizzate da sistemi automatici, adattati e non. L'adattamento è stato eseguito sulla base di una porzione del progetto tradotto durante una fase preliminare, in cui allo stesso traduttore è stato chiesto di correggere le traduzioni fornite da un sistema di partenza non adattato.

[1] http://www.matecat.com

Siccome le valutazioni sul campo sono estremamente costose, esse non possono essere eseguite frequentemente per confrontare tutte le possibili varianti degli algoritmi e dei processi. Abbiamo quindi condotto anche delle *valutazioni di laboratorio*, in cui le correzioni dei traduttori erano simulate dalle traduzioni di riferimento.

Complessivamente, nel dominio legale i miglioramenti osservati in laboratorio hanno anticipato quelli misurati sul campo. Al contrario, i risultati nel dominio TIC sono stati controversi a causa della poca corrispondenza tra i testi usati per l'adattamento e quelli effettivamente tradotti durante la sperimentazione.

2 Lavori correlati

L'idea che la TA possa migliorare la produttività dei traduttori si è consolidata negli anni grazie ai miglioramenti della qualità della TA statistica e ai tanti lavori che hanno sperimentalmente valutato il suo impatto (Guerberof, 2009; Plitt and Masselot, 2010; Federico et al., 2012; Läubli et al., 2013; Green et al., 2013).

Dal punto di vista dei metodi, il nostro lavoro si occupa di adattamento in generale, e di quello incrementale più nello specifico. Senza entrare nel dettaglio per mancanza di spazio, vogliamo qui segnalare il lavoro di Bertoldi et al. (2012), dove i modelli di traduzione vengono adattati incrementalmente su pacchetti di dati nuovi man mano che questi sono disponibili, e i seguenti lavori in qualche modo a quello correlati: (Koehn and Schroeder, 2007; Foster and Kuhn, 2007; Bisazza et al., 2011; Liu et al., 2012; Bach et al., 2009; Niehues and Waibel, 2012; Hasler et al., 2012).

Come vedremo, noi eseguiamo anche una selezione di dati, problema ampiamente investigato dalla nostra comunità scientifica, si veda ad esempio (Yasuda et al., 2008; Matsoukas et al., 2009; Foster et al., 2010; Axelrod et al., 2011). Quella che noi applichiamo è una tecnica piuttosto convenzionale (Moore and Lewis, 2010), ma in condizioni inusuali, dove la selezione viene effettuata su un corpus di dati nel dominio di interesse, e quindi non proprio generico, e a partire da una quantità estremamente ridotta di dati specifici.

3 Metodi di adattamento

Selezione di dati - Come accennato sopra, quello della selezione di dati è un problema ampiamente studiato in letteratura. In effetti spesso ci troviamo a dover addestrare dei modelli con dati provenienti da sorgenti eterogenee in termini di dimensione, qualità, dominio, ecc. L'obiettivo di questa tecnica è di selezionare un sottoinsieme dei dati a disposizione che sia pertinente rispetto ad un certo testo, nel nostro caso quello di uno specifico progetto di traduzione. Noi abbiamo implementato la tecnica proposta in (Moore and Lewis, 2010) e l'abbiamo resa disponibile attraverso il pacchetto IRSTLM (Federico et al., 2008). Per applicare l'algoritmo, si parte da un *corpus specifico*, che si suppone rappresentare bene il documento da tradurre, e da un *corpus generico*, molto più grande e in cui si suppone di poter trovare del materiale pertinente al documento da tradurre. Sfruttando due modelli del linguaggio (ML), uno specifico e uno generico, a ogni frase generica viene assegnato un punteggio tanto più alto quanto più essa è specifica e lontana dalla "media" di quelle generiche. Effettuato l'ordinamento su questo punteggio, viene infine selezionata la quantità di frasi generiche che ottimizzano la perplessità di un testo di controllo.
Fill-up dei modelli di traduzione - La selezione di dati è efficace per addestrare modelli di traduzione sui testi più rilevanti per uno specifico progetto. D'altra parte, scartare una porzione dei dati disponibili significa correre il rischio di perdere delle informazioni comunque utili; per evitarlo, si può ricorrere alla tecnica *fill-up*, proposta da Nakov (2008) e raffinata da Bisazza et al. (2011). Essa fonde i modelli di traduzione generico e specifico, unendo gli insiemi delle loro voci e mantenendo le probabilità del modello specifico per le voci in comune.
Mistura di ML - Per l'adattamento dei ML siamo ricorsi alla mistura dei modelli proposta da Kneser e Steinbiss (1993), che consiste nella combinazione convessa di due o più ML; i pesi della mistura sono stimati per mezzo di una validazione incrociata sui dati di addestramento con la quale si simula l'occorrenza di n-grammi nuovi. Il metodo è disponibile nel già citato pacchetto IRSTLM.

4 Dati per gli esperimenti

Coi domini e le coppie di lingue menzionate nell'introduzione abbiamo definito sei configurazioni sperimentali. Qui di seguito forniamo dettagli sui dati di addestramento e di valutazione per ciascuna di esse.
Dati di addestramento - Per l'addestramnento abbiamo usato sia dati paralleli sia memorie di traduzione. Per il dominio TIC, sono stati sfruttati i manuali software del corpus OPUS (Tiedemann, 2012) e una memoria di traduzione proprietaria, fornitaci dal partner industriale di MateCat.

Per il dominio LGL abbiamo acquisito il corpus JRC-Acquis (Steinberger et al., 2006), che include

dominio	coppia	corpus	seg	parole	
				sorgente	obiettivo
TIC	IT	generico	5.4 M	57.2M	59.9M
		selezione	0.36M	3.8M	4.0M
		calibrazione	2,156	26,080	28,137
	FR	generico	2.3 M	35.4M	40.1M
		selezione	0.53M	8.6M	9.5M
		calibrazione	4,755	26,747	30,100
LGL	IT	generico	2.7 M	61.4M	63.2M
		selezione	0.18M	5.4M	5.4M
		calibrazione	181	5,967	6,510
	FR	generico	2.8 M	65.7M	71.1M
		selezione	0.18M	5.5M	5.8M
		calibrazione	600	17,737	19,613
	ES	generico	2.3 M	56.1M	62.0M
		selezione	0.18M	5.6M	6.1M
		calibrazione	700	32,271	36,748
	DE	generico	2.5 M	45.3M	41.8.0M
		selezione	0.18M	5.2M	4.7M
		calibrazione	133	3,082	3,125

Tabella 1: Statistiche sui dati paralleli usati per la preparazione dei sistemi di TA: numero di segmenti e di parole. Il simbolo M sta per 10^6.

la legislazione della UE in 22 lingue.

La tabella 1 riporta alcune statistiche dei testi paralleli impiegati per l'addestramento dei modelli di traduzione e di riordinamento; i ML sono stati stimati sul testo obiettivo. Per ciascuna configurazione sperimentale, la voce *generico* si riferisce alla totalità dei dati a disposizione, mentre *selezione* indica i dati selezionati pertinenti al progetto in esame. I dati per la *calibrazione* sono aggiuntivi e utilizzati per il bilanciamento ottimale dei vari modelli che definiscono il motore di TA.

Dati di valutazione - Per il dominio TIC i dati sono stati forniti dal partner industriale di MateCat che ha selezionato dal suo archivio un progetto di traduzione reale in cui dei documenti in inglese erano già stati tradotti in italiano e francese senza l'ausilio del sistema MateCat. Per il dominio LGL, abbiamo selezionato un documento della legislazione europea[2] per il quale erano disponibili le traduzioni nelle quattro lingue di nostro interesse. Nella tabella 2 sono raccolte le statistiche dei testi da tradurre nella fase preliminare e in quella di validazione vera e propria del sistema adattato.

5 Valutazioni di laboratorio

Sistemi di TA - I sistemi di TA sviluppati sono statistici e costruiti col pacchetto Moses (Koehn et al., 2007). I modelli di traduzione e di riordinamento sono stati addestrati sui dati bilingue della tabella 1 nei modi descritti in seguito. Per modellare il linguaggio, le distribuzioni dei 5-grammi

[2]2013/488/EU: "Council Decision of 23 September 2013 on the security rules for protecting EU classified information".

dominio	coppia	fase	seg	parole	
				sorgente	obiettivo
TIC	IT	preliminare	342	3,435	3,583
		validazione	1,614	14,388	14,837
	FR	preliminare	342	3,435	3,902
		validazione	1,614	14,388	15,860
LGL	IT	preliminare	133	3,082	3,346
		validazione	472	10,822	11,508
	FR	preliminare	134	3,084	3,695
		validazione	472	10,822	12,810
	ES	preliminare	131	3,007	3,574
		validazione	472	10,822	12,699
	DE	preliminare	133	3,082	3,125
		validazione	472	10,822	10,963

Tabella 2: Statistiche sui dati di valutazione.

sono state stimate sul testo obiettivo e applicandovi la tecnica di smoothing Kneser-Ney (Chen and Goodman, 1999). La calibrazione dei sistemi, ovvero la stima dei pesi dell'interpolazione dei vari modelli, è stata effettuata su opportuni testi aggiuntivi (voci *calibrazione* in tabella 1).

Per ciascuna delle sei configurazioni, sono stati valutati due sistemi TA, uno di riferimento (RIF) e uno adattato (ADA). I modelli del RIF sono stati addestrati sui dati corrispondenti alle voci *generico* della tabella 1. Abbiamo quindi selezionato una porzione dei dati generici usando come corpus specifico il testo bilingue della fase preliminare e la parte sorgente del testo di validazione, ottenendo i testi *selezione* della tabella 1. Sulla concatenazione dei testi della fase preliminare e di quelli selezionati abbiamo successivamente addestrato i modelli specifici che sono stati combinati coi modelli generici per mezzo del fill-up (modelli di traduzione/riordinamento) e della mistura (ML) al fine di costruire il sistema ADA.

Risultati - La tabella 3 quantifica la qualità della TA fornita dai sistemi RIF e ADA in termini di Bleu, Ter e Gtm, misurati sui documenti di validazione rispetto alle traduzioni manuali.

coppia	TA	dominio TIC			dominio LGL		
		Bleu	Ter	Gtm	Bleu	Ter	Gtm
IT	RIF	55.3	29.2	77.8	31.0	53.1	61.8
	ADA	57.5	26.3	78.6	35.0	49.1	64.6
FR	RIF	41.3	38.3	69.5	33.9	52.2	63.0
	ADA	41.4	37.9	69.9	36.4	49.1	65.1
ES	RIF	–	–	–	35.5	50.7	65.7
	ADA	–	–	–	36.4	50.2	65.6
DE	RIF	–	–	–	18.3	68.4	50.5
	ADA	–	–	–	19.7	66.6	52.3

Tabella 3: Prestazioni TA sui testi di validazione .

Nel dominio LGL il miglioramento fornito dal processo di adattamento è rilevante. Ad esempio, il Bleu migliora del 12.9% (da 31.0 a 35.0) nella traduzione in italiano, del 7.4% (da 33.9 a 36.4)

verso il francese, del 2.5% (da 35.5 a 36.4) verso lo spagnolo e del 7.7% (da 18.3 a 19.7) verso il tedesco.

Al contrario, nel TIC si osserva un certo miglioramento solo per l'italiano (4%, da 55.3 a 57.5), mentre è nullo per il francese. L'analisi riportata in (Bertoldi et al., 2013) mostra che qui il problema è originato dal fatto che i testi tradotti nella fase preliminare, e quindi usati per la selezione, sono poco rappresentativi del documento da tradurre nella fase di validazione.

6 Valutazioni sul campo

In questo paragrafo relazioniamo sugli esperimenti effettuati per valutare l'impatto dell'adattamento al progetto sulla produttività di traduttori professionisti. La valutazione sul campo ha riguardato la traduzione dall'inglese all'italiano di documenti nei due domini TIC e LGL.

Protocollo - La valutazione sul campo è stata eseguita con il sistema di ausilio alla traduzione sviluppato nell'ambito del progetto MateCat che integra i sistemi di TA auto-adattantisi al progetto, come descritto in questo articolo. L'esperimento è stato organizzato su due giorni ed ha coinvolto quattro traduttori per ciascun dominio. Durante il primo giorno – la fase preliminare – per la traduzione della prima parte del progetto i suggerimenti di TA venivano forniti dal sistema RIF; nel secondo giorno – la fase di validazione –, durante il quale è stata tradotta la seconda parte del progetto, i suggerimenti di TA provenivano dal sistema ADA. L'impatto dello schema di adattamento proposto in questo articolo è stato misurato confrontando la produttività dello stesso traduttore nel primo e nel secondo giorno, misurata in termini di time-to-edit (TTE)[3] e post-editing effort (PEE).[3]

Risultati - I risultati sono raccolti nella tabella 4. Per due traduttori su quattro nel dominio TIC (t1 e t4) e per tre su quattro nel LGL (t2-t4) migliorano significativamente entrambe le misure. La maggior parte delle riduzioni del TTE (cinque su otto) sono statisticamente significative (p-value$<$0.05), mentre lo stesso accade solo per due delle variazioni del PEE. Guardando alle medie, nel dominio TIC si registra un guadagno dell'11.2% del TTE e del 6.5% del PEE, mentre nel LGL i miglioramenti sono rispettivamente del 22.2% e del 10.7%. Infine, la buona correlazione osservata tra PEE e TTE nelle diverse condizioni sperimentate mostra come sia verosimile che i traduttori abbiano tratto beneficio dai suggerimenti provenienti dal sistema di TA adattato, dato che il PEE è migliorato in sette casi su otto.

[3]In breve, il TTE è il tempo medio (in secondi) di traduzione per parola, il PEE la percentuale di parole che sono state corrette.

mtrc	dmn	usr	prlmnr	vldzn	p-value	Δ
TTE	TIC	t1	4.70	3.36	0.001	28.51%
		t2	2.26	2.47	0.220	-9.29%
		t3	3.17	3.11	0.450	1.89%
		t4	4.77	3.64	0.006	23.69%
	LGL	t1	5.20	5.63	0.222	-8.27%
		t2	5.42	3.92	0.002	27.68%
		t3	5.86	4.32	0.000	26.28%
		t4	6.60	3.73	0.000	43.48%
PEE	TIC	t1	34.27	30.99	0.060	9.57%
		t2	38.50	39.52	0.330	-2.65%
		t3	32.53	30.17	0.133	7.25%
		t4	32.22	28.44	0.040	11.73%
	LGL	t1	26.47	24.57	0.212	7.18%
		t2	29.11	26.25	0.140	9.82%
		t3	35.65	34.11	0.247	4.32%
		t4	22.72	18.07	0.011	20.47%

Tabella 4: TTE e PEE di ciascun traduttore nelle due sessioni, la preliminare (*prlmnr*) e di validazione (*vldzn*). Sono riportate anche la differenza dei valori tra le due sessioni e la sua significatività statistica in termini di p-value, calcolato tramite la versione randomizzata del test di permutazione (Noreen, 1989).

7 Conclusioni

Un argomento di ricerca particolarmente attuale per l'industria della traduzione assistita è come dotare i sistemi di traduzione automatica della capacità di auto-adattamento. In questo lavoro abbiamo presentato uno schema di auto-adattamento ed i risultati della sua validazione non solo in esperimenti di laboratorio ma anche sul campo, col coinvogimento di traduttori professionisti, grazie alla collaborazione col partner industriale di MateCat.

I risultati sperimentali hanno confermato l'efficacia della nostra proposta, essendosi ottenuti guadagni di produttività fino al 43%. Tuttavia, il metodo funziona solo se i testi utilizzati come base per la selezione di dati specifici su cui eseguire l'adattamento è rappresentativo del documento che si vuol far tradurre. Infatti, laddove tale condizione non fosse verificata, com'era nei nostri esperimenti inglese-francese/TIC, i modelli adattati possono risultare incapaci di migliorare quelli di partenza; ad ogni modo anche in queste condizioni critiche non abbiamo osservato alcun deterioramento delle prestazioni, a dimostrazione del comportamento conservativo del nostro schema.

Ringraziamenti

Questo lavoro è stato possibile grazie al progetto MateCat, finanziato dalla Commissione Europea nell'ambito del Settimo programma quadro.

References

A. Axelrod, X. He, and J. Gao. 2011. Domain Adaptation via Pseudo In-Domain Data Selection. In *EMNLP*, pp. 355–362, Edinburgh, UK.

N. Bach, R. Hsiao, M. Eck, P. Charoenpornsawat, S. Vogel, T. Schultz, I. Lane, A. Waibel, and A. W. Black. 2009. Incremental Adaptation of Speech-to-Speech Translation. In *NAACL HLT (Short Papers)*, pp. 149–152, Boulder, US-CO.

N. Bertoldi, M. Cettolo, M. Federico, and C. Buck. 2012. Evaluating the Learning Curve of Domain Adaptive Statistical MachineTranslation Systems. In *WMT*, pp 433–441, Montréal, Canada.

N. Bertoldi, M. Cettolo, M. Negri, M. Turchi, M. Federico, and H. Schwenk. 2013. D5.4: Second report on lab and field test. Deliverable, MateCat project. http://www.matecat.com/wp-content/uploads/2014/06/D5.4_Second-Report-on-Lab-and-Field-Test_v2.pdf.

A. Bisazza, N. Ruiz, and M. Federico. 2011. Fill-up versus Interpolation Methods for Phrase-based SMT Adaptation. In *IWSLT*, pp. 136–143, San Francisco, US-CA.

S. F. Chen and J. Goodman. 1999. An Empirical Study of Smoothing Techniques for Language Modeling. *Computer Speech and Language*, 4(13):359–393.

M. Federico, N. Bertoldi, and M. Cettolo. 2008. IRSTLM: an Open Source Toolkit for Handling Large Scale Language Models. In *Interspeech*, pp. 1618–1621, Melbourne, Australia.

M. Federico, A. Cattelan, and M. Trombetti. 2012. Measuring user productivity in machine translation enhanced computer assisted translation. In *AMTA*, San Diego, US-CA.

G. Foster and R. Kuhn. 2007. Mixture-model Adaptation for SMT. In *WMT*, pp. 128–135, Prague, Czech Republic.

G. Foster, C. Goutte, and R. Kuhn. 2010. Discriminative Instance Weighting for Domain Adaptation in Statistical Machine Translation. In *EMNLP*, pp. 451–459, Cambridge, US-MA.

S. Green, J. Heer, and C. D Manning. 2013. The efficacy of human post-editing for language translation. In *SIGCHI Conference on Human Factors in Computing Systems*, pp. 439–448, Paris, France.

A. Guerberof. 2009. Productivity and quality in MT post-editing. In *MT Summit - Beyond Translation Memories: New Tools for Translators Workshop*.

E. Hasler, B. Haddow, and P. Koehn. 2012. Sparse lexicalised features and topic adaptation for SMT. In *IWSLT*, pp. 268–275, Hong-Kong (China).

R. Kneser and V. Steinbiss. 1993. On the dynamic adaptation of stochastic language models. In *ICASSP*, volume II, pp. 586–588, Minneapolis, US-MN.

P. Koehn and J. Schroeder. 2007. Experiments in Domain Adaptation for Statistical Machine Translation. In *WMT*, pp. 224–227, Prague, Czech Republic.

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *ACL: Companion Volume Proceedings of the Demo and Poster Sessions*, pp. 177–180, Prague, Czech Republic.

S. Läubli, M. Fishel, G. Massey, M. Ehrensberger-Dow, and M. Volk. 2013. Assessing Post-Editing Efficiency in a Realistic Translation Environment. In *MT Summit Workshop on Post-editing Technology and Practice*, pp 83–91, Nice, France.

L. Liu, H. Cao, T. Watanabe, T. Zhao, M. Yu, and C. Zhu. 2012. Locally Training the Log-Linear Model for SMT. In *EMNLP-CoNLL*, pp. 402–411, Jeju Island, Korea.

S. Matsoukas, A.-V. I. Rosti, and B. Zhang. 2009. Discriminative Corpus Weight Estimation for Machine Translation. In *EMNLP*, pp. 708–717, Singapore.

R. C. Moore and W. Lewis. 2010. Intelligent Selection of Language Model Training Data. In *ACL (Short Papers)*, pp. 220–224.

P. Nakov. 2008. Improving English-Spanish Statistical Machine Translation: Experiments in Domain Adaptation, Sentence Paraphrasing, Tokenization, and Recasing. In *WMT*, pp. 147–150, Columbus, US-OH.

J. Niehues and A. Waibel. 2012. Detailed Analysis of different Strategies for Phrase Table Adaptation in SMT. In *AMTA*, San Diego, US-CA.

E. W. Noreen. 1989. *Computer Intensive Methods for Testing Hypotheses: An Introduction*. Wiley Interscience.

M. Plitt and F. Masselot. 2010. A Productivity Test of Statistical Machine Translation Post-Editing in a Typical Localisation Context. *Prague Bulletin of Mathematical Linguistics*, 93:7–16.

R. Steinberger, B. Pouliquen, A. Widiger, C. Ignat, T. Erjavec, D. Tufi, and D. Varga. 2006. The JRC-Acquis: A Multilingual Aligned Parallel Corpus with 20+ Languages. In *LREC*, pp. 2142–2147, Genoa, Italy.

J. Tiedemann. 2012. Parallel Data, Tools and Interfaces in OPUS. In *LREC*, Istanbul, Turkey.

K. Yasuda, R. Zhang, H. Yamamoto, and E. Sumita. 2008. Method of Selecting Training Data to Build a Compact and Efficient Translation Model. In *IJCNLP*, Hyderabad, India.