Utvärderingsansatser för MATS-systemet

Eva Forsbom

Institutionen för lingvistik

Uppsala universitet

2002-08-22

Korpusöversikt

Den korpus av parallellställda texter vi utgår från är uppdelad på en träningsdel och en utvärderingsdel. De svenska texterna har följande fördelning:

	Träning		Utv.	
	Totalt	Unika	Totalt	Unika
Ord	44196	6842	11194	2392
Segment	7412	4734	1772	1187

Tabell 1: Vad som ska översättas 2002-05-20

Utvärdering av systemets fullständighet

Genom att lägga på en utvärderingsmodul i systemet kan vi räkna fram hur många segment som översätts, helt och hållet eller med eventuellt oöversatta ord.

	Träning		Utv.	
	Totalt	Unika	Totalt	Unika
Översatta	41.3	28.4	39.3	25.9
Helt översatta	36.9	22.4	34.5	19.0

Tabell 2: Översatta segment 2002-05-20

Utvärdering av modulernas felrapporter 1

Lexikon

Utvärderingsmodulen samlar även ihop och summerar data och felrapporter från de olika modulerna.

	Träning		Utv.	
	Totalt	Unika	Totalt	Unika
Källspråk	193	88	112	52
Översättning[a]	13075	1900	2740	607
Målspråk[b]	925	315	174	9

Tabell 3: Ej funna ord 2002-05-20

[a] Ej betydelsetolkade

[b] Inkluderar kopierade källspråksord

Utvärdering av modulernas felrapporter 2

Grammatiker

		Träning		Utv.
	Totalt	Unika	Totalt	Unika
Delvis parsade	1473	1040	355	285
Ej parsade[a]	2560	2275	690	581
Ej överförda	266	27	18	8
Ej genererade	53	46	12	6
Fel mosykod	0	0	0	0
Ingen felrapport	2737	1062	612	225

Tabell 4: (Ej) klarade segment 2002-05-20

[a] Fel i grammatiken

Utvärdering av modulernas felrapporter 3

Loggar

Utvärderingsmodulen kan också generera loggar över valfria fel. Loggarna kan sedan användas för uppdatering av lexikon och grammatiker.

Exempel

- Vi har använt loggarna över ord som saknades i källspråkslexikonet till att förbättra användningen av reguljära uttryck för exempelvis alfanumeriska typbeteckningar i lexikonen.
- Vi har vänt på systemet och använt loggarna över ord som saknades i det källspråkslexikonet (engelska) till att uppdatera det engelska lexikonet i den aktuella översättningsriktningen.
- Vi har under sommaren gått igenom en logg över de 1 900 ord som saknade översättningar och tagit fram översättningarna till dem.
- Vi kikar just nu på lämpliga sätt att använda loggarna från parsern till att effektivisera grammatikskrivandet.

Utvärdering av systemets noggrannhet 1

Mått

Under sommaren har vi börjat undersöka ett par automatiska sätt att mäta hur pass väl systemets översättningar stämmer överens med korpusens facit.

- BLEU/NIST[a] – som är baserat på jämförelser av n-gram.
- Ordnoggrannhet (ON)[b] – som är baserat på antalet redigeringar mellan systemets version och facit.

[a] Papineni et al. (2001)
[b] Se t.ex. Alshawi et al. (1998)

Utvärdering av systemets noggrannhet 2

Måtten BLEU och NIST

BLEU/NIST har använts som mått vid DARPA för utvärderingar av maskinöversättningssystem, och visat sig korrelera väl med mänskliga utvärderares bedömningar. De har främst testats på engelska och kinesiska, och framför allt på nyhetstexter, ofta med flera facit åt gången. NIST är en utveckling av BLEU.

BLEU

- Beräknar summan av alla 1-gram, 2-gram, 3-gram och 4-gram som förekommer i både maskinöversättningen och facit i förhållande till segmentets längd.
- Mäter om orden är rätt översatta (1-gram) och om texten är idiomatisk.
- Mäts i intervallet 0–1.

NIST

- Använder en annan funktion som kompenserar för långa segment.
- Mäts i intervallet 0–∞.
- Tolkas i förhållande till andra system.

Utvärdering av systemets noggrannhet 3

Måttet ON

ON (och redigeringsavstånd) har använts som mått för utvärderingar av ett flertal tillämpningar för naturligt språk. Vi håller på att ta fram ett program som använder samma indataformat och förbehandling av materialet som i BLEU/NIST-programmet (t.ex. konvertering av versal till gemen).

- Beräknar det minsta antalet redigeringar som behövs för att maskinöversättningen ska stämma med facit, i förhållande till segmentets längd, och drar det från idealet 1.
- $1 - \frac{I+R+E}{L}$
- Redigeringarna är insättning, radering och ersättning.
- Mäter om orden är rätt översatta och om texten är idiomatisk.
- Mäts i intervallet 0–1.

Utvärdering av systemets noggrannhet 4

Inledande experiment

Vi har använt måtten BLEU och ON på 1 873 segment som översatts i systemet. Det publika programmet för BLEU/NIST ger felaktiga siffror för segment med färre än 4 ord (1 660 st.), men för övriga segment verkar siffrorna vettiga.

	BLEU orig.	BLEU rev.[a]	ON
Systemnivå	52	52	85
Segmentsnitt (alla)	4	94	85
Segmentsnitt (>4)	32	71	64

Tabell 5: Noggrannhet, i % (2002-06-18).

[a]Språkspecifika ändringar och försök till korrigering

Utvärdering av systemets noggrannhet 5

Exempel på skillnader

De båda måtten är känsliga för olika fenomen, som fel på ordnivå och fel på ordföljd.

Fel på ordnivå: numerus

Källa slang mot termostathus

System hose to thermostat housings

Facit hose to thermostat housing

ON 75 %

BLEU 71 %

Fel på ordföljd

Källa magnetventiler för insprutningstidpunkt

System solenoid valves for injection timing

Facit injection timing solenoid valves

ON 0 %

BLEU 80 %

Referenser

- Alshawi, H, Bangalore, S., & Douglas, S., 1998. Automatic Acquisition of Hierarchical Transduction Models for Machine Translation. I *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics*, s. 41–47, Montreal, Canada.
- Papineni, K., Roukos, S., Ward, T., & Zhu, W-J, 2001. *Bleu: A Method for Automatic Evaluation of Machine Translation.* IBM Report RC22176.
- http://www.nist.gov/speech/tests/mt/