Evaluación de un sistema de traducción automática basado en reglas

o por qué BLEU sólo sirve para lo que sirve

Aingeru Mayor, Iñaki Alegria, Arantza Díaz de Ilarraza, Gorka Labaka, Mikel Lersundi, Kepa Sarasola

IXA Taldea
Euskal Herriko Unibertsitatea
aingeru@ehu.es

2009ko irailaren 8a

Introducción

- Evaluación del sistema de TA *Matxin* es→eu
  - Absoluta
    - Medida total del comportamiento del sistema
  - Relativa
    - Comparar con el sistema basado en corpus *Matrex*
- Métricas usadas
  - HTER
    - Requiere postedición manual
    - Proporciona una medida significativa
  - BLEU
    - Métrica automática
    - Conclusiones sobre su uso

Índice

Índice

Matxin

- Primer sistema de TA públicamente disponible que traduce a euskera (Mayor, 07; Alegría *et al.*,08).
- Desarrollado por el grupo IXA
- Basado en reglas (modelo tradicional de transferencia)
- Software de código abierto
- *Matxin1.0*: es→eu
  - http://www.opentrad.org
  - http://www.matxin.sourceforge.net

Matrex

- Sistema basado en corpus
- Adaptacíon es⟶eu (Way *et al.*, 06; Labaka *et al.*, 07)
  - Usando herramientas de procesamiento del euskera
  - Entrenado con 50.000 oraciones del corpus *Consumer*

Índice

Evaluación en TA

- La pregunta correcta quizá no sea cuánto de buena es la traducción automática, sino cuán utilizable es (Koehn *et al.*, 07).

- Tipos de evaluación:
  - manual
    - significativa
    - cara y no reutilizable
  - automática
    - rápida, repetible y objetiva
    - no garantiza resultados correctos

Evaluación manual

- Medidas usadas tradicionalmente
  - fidelidad (*adequacy*)
  - fluidez (*fluency*)
- Una alternativa: HTER (Snover *et al.*, 06)
  - *Human-targeted Translation Edit Rate*
  - También llamada distancia de edición (Przybocki *et al.*, 06)
  - Calcula el coste de postedición de la traducción dada por el sistema

HTER

- Un editor humano modifica la traducción del sistema de TA
  - de manera que la versión editada
    - tenga *el mismo significado* del texto de origen
    - sea *comprensible* y *gramaticalmente correcta*
  - realizando el mínimo número de modificaciones posible
    - inserción, borrado y sustitución de palabras
    - movimiento de grupos de palabras

$$HTER = \frac{num\_modificaciones}{num\_tokens\_traduccion\_editada}$$

HTER

- Muestra cuánto de utilizables son las traducciones obtenidas
- Mejor correlación que BLEU con los juicios humanos de fidelidad y fluidez (Snover *et al.*, 06; Przybocki *et al.*, 06)

- Principal problema: su coste (500 palabras/hora)
  - Asumible: para evaluar uno o unos pocos sistemas
  - No asumible: en campañas de evaluación con muchos sistemas

Evaluación automática

- Se compara la traducción del sistema con traducciones humanas de referencia
- Hipótesis: cuanto más parecidas sean, mejor será la traducción.
- Necesario usar un conjunto de traducciones de referencia, si bien en la mayoría de los casos sólo se tiene disponible una

Evaluación automática

- Cálculo de la proximidad entre la traducción del sistema y las de referencia
  - basándose en *string matching*
    - WER (Nieben *et al.*, 00)
    - PER (Leusch *et al.*, 03)
    - TER (Snover *et al.*, 06)
  - basándose en *n*-gramas
    - BLEU (Papineni *et al.*, 02)
    - NIST (Doddington *et al.*, 02)
    - WNM (Babych *et al.*, 04)
    - F-measure (Melamed *et al.*, 03)
    - Meteor (Lavie *et al.*, 04)

BLEU: luces...

- Medida de evaluación de TA más usada hoy en día
  - Ha guiado el progreso en el desarrollo de los sistemas estadísticos
  - Medida elegida para campañas de evaluación como NIST
- (Papineni *et al.*, 02; Doddington *et al.*, 02) afirman que:
  - rápida, barata, independiente del lenguaje
  - gran correlación con las evaluaciones manuales

BLEU: ...y sombras

- Difícil interpretar lo que expresa un resultado BLEU
- (Callison-Burch *et al.*, 06, Koehn *et al.*, 06) han demostrado que
  - Mejora de BLEU $\not\Leftrightarrow$ Mejora en la calidad de la traducción
  - No tiene una correlación tan alta como se cree
    - En la campaña de evaluación NIST 2005:
      1° en la evaluación manual, se clasificó como 6° con BLEU
    - BLEU beneficia a los sistemas estadísticos

BLEU: ...y sombras

- (Boitet *et al.*, 06) Las métricas basadas en *n*-gramas son
  - Inadecuadas
    - No miden la calidad de la traducción,
      sino su parecido con las traducciones de referencia
  - Caras
    - El coste de preparar las traducciones de referencia es muy alto

- (Homola *et al.*, 09) Resultados aún más inadecuados para
  - lenguas con una morfología más rica
  - o con mayor grado de libertad en el orden de las palabras

BELU: Usos adecuados e inadecuados

(Callison-Burch *et al.*, 06) Es necesario distinguir

- Usos adecuados
  - Seguimiento de los cambios incrementales de un sistema
  - Comparación de sistemas que usen estrategias similares
  - Optimización de los valores de los parámetros de sistemas estadísticos

BELU: Usos adecuados e inadecuados

- Usos inadecuados
  - Comparación de sistemas
    - que usen estrategias diferentes
    - cuando el par de lenguas, el número de referencias o el tamaño de *n*-gramas es diferente
  - Identificación de mejoras de aspectos de la traducción que la métrica no modela bien
  - Monitorización de mejoras que aparecen poco en el corpus de test

BLEU: Situación actual

- Excesiva confianza hacia BLEU
- En las pubilcaciones científicas
  - Se presentan mejoras en la calidad de sistemas de TA, dando únicamente mejores resultados de BLEU sin mostrar ni un solo ejemplo de traducción
  - Se comparan sistemas usando BLEU sin contrastar con evaluaciones manuales
- En la revisión de artículos para su aceptación en congresos
  - Se está pidiendo que se incluyan resultados BLEU, en situaciones en las que su uso es totalmente inadecuado

Índice

Diseño de la evaluación

- Corpus
  - *Eitb*, corpus periodístico general
  - *Consumer*, corpus sobre consumo
- Oraciones de entre 5 y 25 palabras elegidas al azar
- HTER
  - 50 oraciones de cada corpus
  - Editor bilingüe
- BLEU
  - 1.500 oraciones de cada corpus
  - Una sola traducción de referencia

Resultados HTER

Matxin

	Ins	Bor	Sus	Mov	Ed	Tok	HTER	
Eitb	20	13	147	47	215	532	40,41%	42,00%
Cons	27	30	152	56	259	594	43,60%	

Matrex

	Ins	Bor	Sus	Mov	Ed	Tok	HTER	
Eitb	35	101	205	27	368	512	71,87%	64,92%
Cons	47	55	187	60	349	602	57,97%	

* Valores HTER más pequeños indican mejor calidad

Resultados HTER

	HTER	
	Matxin	*Matrex*
Eitb	**40,41%**	71,87%
Consumer	**43,60%**	57,97%

* Valores HTER más pequeños indican mejor calidad

- Evaluación absoluta
  - *Matxin*: 4 de cada 10 palabras han de ser posteditadas
  - *Matrex*: 6-7 de cada 10 palabras han de ser posteditadas
- Evaluación relativa
  - *Matxin* significativamente mejor que *Matrex*

Resultados BLEU

	BLEU	
	Matxin	*Matrex*
Eitb	**9,30**	9,02
Consumer	6,31	**8,03**

* Valores BLEU más grandes deberían indican mejor calidad

- Evaluación absoluta
  - Ninguna información
- Evaluación relativa
  - *Matxin* mejor resultado para el corpus *Eitb*
  - *Matrex* mejor resultado para el corpus *Consumer* !!!

HTER vs BLEU

	HTER		*BLEU*	
	Matxin	*Matrex*	*Matxin*	*Matrex*
Eitb	**40,41%**	71,87%	**9,30**	9,02
Consumer	**43,60%**	57,97%	6,31	**8,03**

- Los resultados HTER basados en la postedición manual invalidan las conclusiones basadas en los resultados BLEU

Índice

Matxin

- Un editor necesitaría cambiar 4 de cada 10 palabras para corregir la salida del sistema
- Mucho que mejorar
- No sirve como sistema para diseminación

- Traducir de español a euskera es una tarea compleja
- Único sistema de TA públicamente disponible que traduce a euskera

HTER

- Evaluación significativa
- Coste no excesivo
  - 200 oraciones
  - 7 horas
  - 350 palabras/hora

BLEU

- No sirve para
  - evaluación de la calidad del sistema *Matxin*
  - comparar el sitema basado en reglas *Matxin* con el sistema basado en corpus *Matrex*
- Herramienta esencial en la construcción de sistemas de TA estadística
- Se usa (y se exige su uso) en situaciones para las que no es apropiada

BLEU

Premisa equivocada

Usamos BLEU, suponiendo que nos sirve, porque es barato

Dos grandes errores:

BLEU

Premisa equivocada

Usamos BLEU, suponiendo que nos sirve, porque es barato

Dos grandes errores:

- BLEU no sirve para lo que no sirve
  - Ofrecer información sobre la calidad absoluta
  - Mostrar en qué medida son utilizables las traducciones realizadas
  - Comparar sistemas que usan estrategias diferentes

BLEU

Premisa equivocada

Usamos BLEU, suponiendo que nos sirve, porque es barato

Dos grandes errores:

- BLEU no sirve para lo que no sirve
  - Ofrecer información sobre la calidad absoluta
  - Mostrar en qué medida son utilizables las traducciones realizadas
  - Comparar sistemas que usan estrategias diferentes
- BLEU sí tiene coste.
  - Es necesario un conjunto detraducciones de referencia, y normalmente solo se tiene una.
  - Crear manualmente las referencias es muy caro

BLEU

- Es urgente
  - dejar a un lado los usos inapropiados de BLEU
  - usar esta métrica
    - sólo cuando realmente sea adecuada
    - y analizando su coste.

Índice

Trabajo futuro

- Aumentar el corpus de evaluación
- Desarrollo de un entorno gráfico de postedición
  - Reutilización las traducciones ya posteditadas en evaluaciones anteriores.
- Cálculo de HTER mejorado (Snover *et al.*, 09)
  - Usa morfología, sinónimos y paráfrasis
  - Ajusta los costes para diferentes tipos de errores
- Comparar cualitativamente las traducciones de los sistemas *Matxin* y *Matrex*
  - En qué acierta y en qué falla cada uno

Evaluación de un sistema de traducción automática basado en reglas

o por qué BLEU sólo sirve para lo que sirve

Aingeru Mayor, Iñaki Alegria, Arantza Díaz de Ilarraza, Gorka Labaka, Mikel Lersundi, Kepa Sarasola

IXA Taldea
Euskal Herriko Unibertsitatea
aingeru@ehu.es

2009ko irailaren 8a