Προσαρμογή Γνωστικού Αντικειμένου σε Συστήματα Στατιστικής Μηχανικής Μετάφρασης

(MSc in AI - UoE 2007)

Δημήτρης Μαυροειδής

Μηχανική Μετάφραση (Ιστορικά στοιχεία 1)

- 1668, Bishop John Wilkins
  - Μεγάλη Οντολογία
- 1933, Petr Smirnov-Troyanskii
  - Μοντέλο Μηχανικής Μετάφρασης με 3 βήματα.
- 1949, Warren Weaver's memorandum
  - «Ένα βιβλίο γραμμένο στα κινέζικα είναι απλά ένα βιβλίο γραμμένο στα αγγλικά 'κωδικοποιημένο' στον 'Κινεζικό κώδικα'. Έχοντας τη δυνατότητα να λύσουμε τα περισσότερα προβλήματα κρυπτογραφίας, δεν έχουμε ήδη χρήσιμες μεθόδους για μετάφραση; […]"

Μηχανική Μετάφραση (Ιστορικά στοιχεία 2)

- 1954, Το πρώτο σύστημα ΜΜ από την IBM
  - Ρωσικά σε αγγλικά
  - Λεξιλόγιο 250 λέξεων
- 1966, ALPAC
  - «Δεν υπάρχει άμεση ή προβλέψιμη προοπτική χρήσιμης Μηχανικής Μετάφρασης»
- 1976, Σύστημα METEO (Καναδάς)
- 1976, SYSTRAN (Ευρωπαϊκή Επιτροπή)

Προσεγγίσεις στη ΜΜ

- Βασισμένη σε Λεξικά (Dictionary-based)
- Βασισμένη σε Κανόνες (Rule-based)
- Βασισμένη σε Παραδείγματα (Example-based)
- **Στατιστική (Statistical)**
- Υβριδική (Hybrid)

Στατιστική Μηχανική Μετάφραση (1)

- Μηχανική Μάθηση (Machine Learning)
- Κανόνας του Bayes:

$$P(T \mid S) = \frac{P(T)P(S \mid T)}{P(S)}$$

$$\arg\max_T P(T \mid S) = \arg\max_T P(T)P(S \mid T)$$

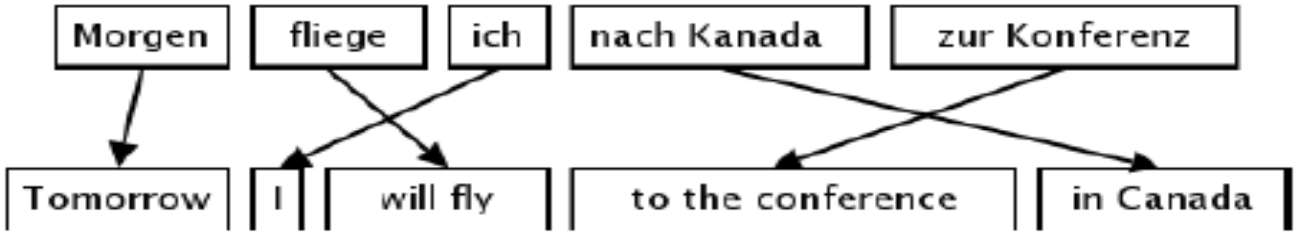
S = Γλώσσα πηγή

T = Γλώσσα στόχος

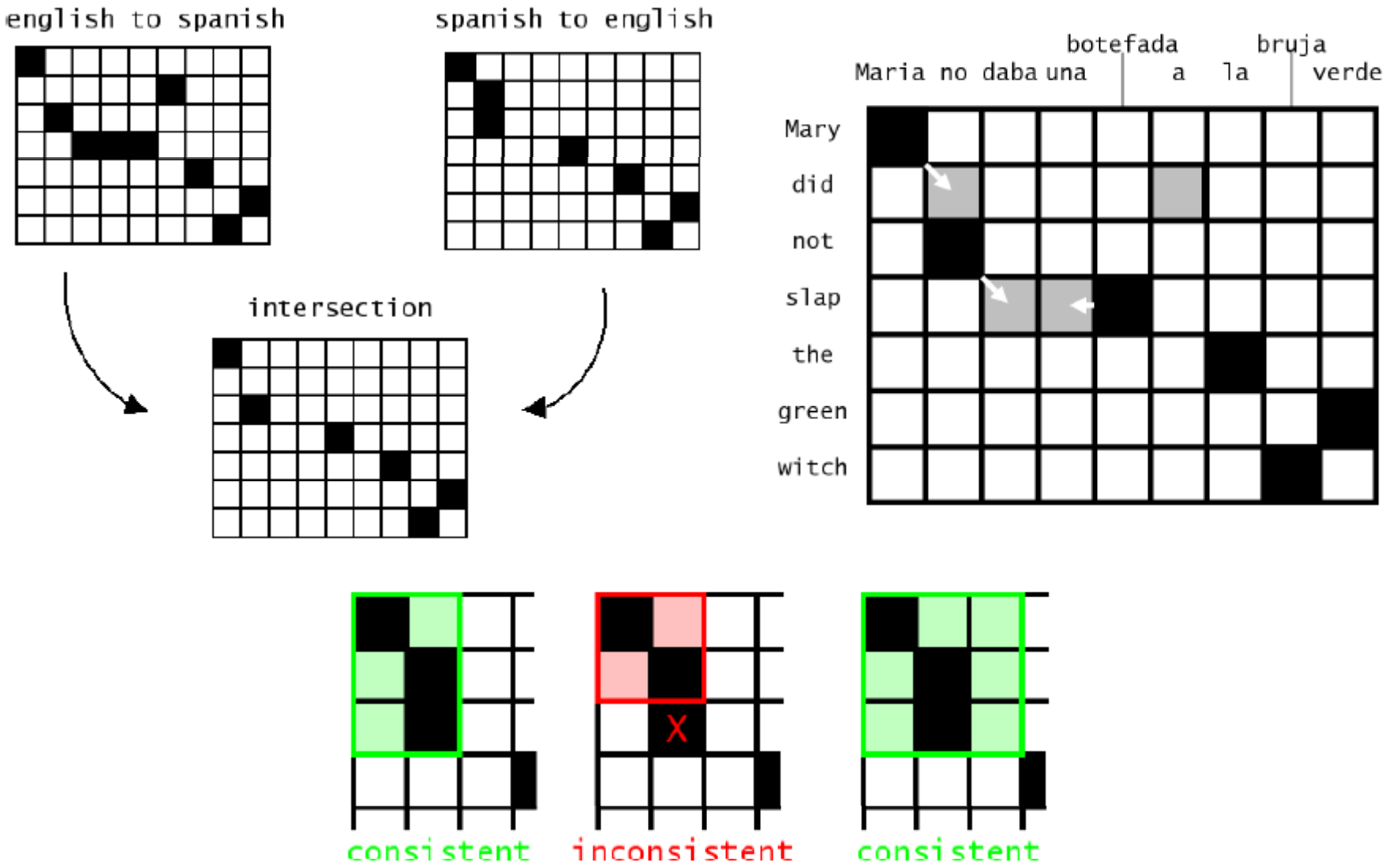
Στατιστική Μηχανική Μετάφραση (2)

- Δίγλωσσο σώμα εκπαίδευσης
- IBM (5 διαδοχικά μοντέλα)
- Μοντέλα στηριγμένα σε λέξεις (word-based)
- Μοντέλα στηριγμένα σε φράσεις (phrase-based):

$$\arg\max_T P(T \mid S) = \arg\max_T P(T)P(S \mid T)\omega^{\text{length(T)}}$$

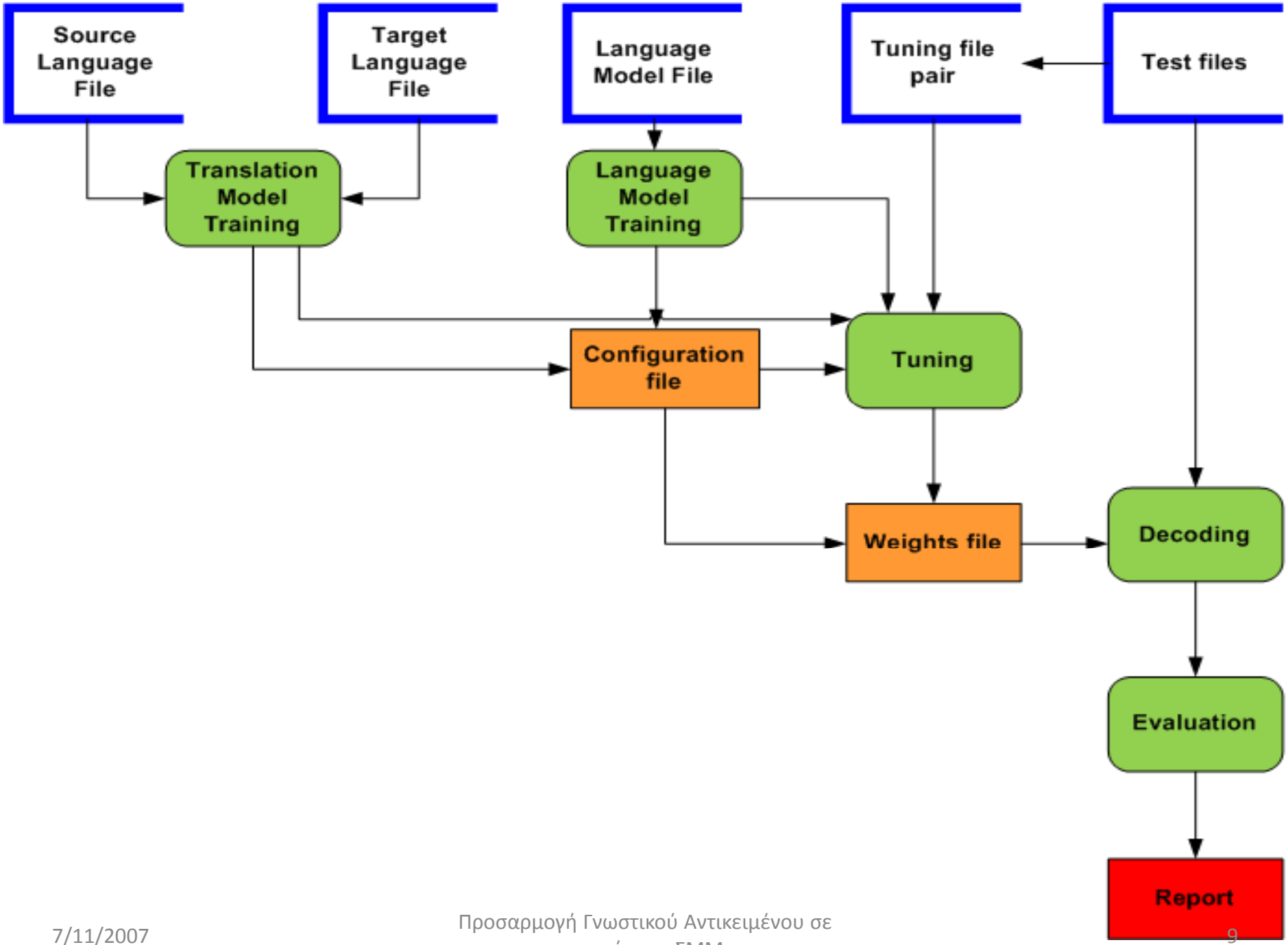


Ευθυγράμμιση



Λειτουργία ενός συστήματος ΣΜΜ

- **Προεπεξεργασία δεδομένων (Pre-processing)**
- **Εκπαίδευση μοντέλων (Models training)**
  - Μοντέλο Μετάφρασης (ΜΜ)
  - Μοντέλο Γλώσσας (ΜΓ)
- **Υπολογισμός βέλτιστων βαρών (Tuning)**
- **Αποκωδικοποίηση (Decoding)**
- **Αξιολόγηση (Evaluation)**
  - Επαγγελματίες μεταφραστές, BLEU, NIST, METEOR, WER


Source Language File
Target Language File
Language Model File
Tuning file pair
Test files
Translation Model Training
Language Model Training
Configuration file
Tuning
Weights file
Decoding
Evaluation
Report

Προεπεξεργασία δεδομένων

Αριθμός Γραμμής	Αγγλικό αρχείο	Ελληνικό αρχείο
1	COMMISSION REGULATION (EEC) No 3812/85	ΚΑΝΟΝΙΣΜΟΣ (ΕΟΚ) αριθ. 3812/85 ΤΗΣ ΕΠΙΤΡΟΠΗΣ
2	of 20 December 1985	της 20ής Δεκεμβρίου 1985
3	adjusting certain Regulations on milk and milk products by reason of the accession of Spain	για την αναπροσαρμογή ορισμένων κανονισμών στον τομέα του γάλακτος και των γαλακτοκομικών προϊόντων, λόγω της προσχώρησης της Ισπανίας
4	THE COMMISSION OF THE EUROPEAN COMMUNITIES,	Η ΕΠΙΤΡΟΠΗ ΤΩΝ ΕΥΡΩΠΑΪΚΩΝ ΚΟΙΝΟΤΗΤΩΝ,
5	Having regard to the Treaty establishing the European Economic Community,	Έχοντας υπόψη: τη συνθήκη για την ίδρυση της Ευρωπαϊκής Οικονομικής Κοινότητας,
6	Having regard to the Act of Accession of Spain and Portugal	την πράξη προσχώρησης της Ισπανίας και της Πορτογαλίας

Εκπαίδευση Μοντέλου Γλώσσας

```
-0.6557779  for community financing , and
-0.3547479  sources of financing , </s>
-0.190566   investigation , findings , affirmative
-0.3547479  and epidemiological findings , a
-0.1818761  or final findings , affirmative
-0.1103589  the provisional findings , as
-0.8318691  the commission finds , following
-0.8318691  the commission finds , on
-0.5308391  of the fine , regard
-0.30103    austria , finland , iceland
-0.4393327  austria , finland , norway
-1.396141   austria , finland , sweden
```

Εκπαίδευση Μοντέλου Μετάφρασης

```
1  του καρδιαγγειακού κινδύνου ||| cardiovascular disease ||| () (0,1) ()
   ||| (1) (1) ||| 0.0212766 1.03513e-05 0.333333 0.0658436 2.718
2  του καρδιαγγειακού ||| cardiovascular disease ||| () (0,1) ||| (1) (1)
   ||| 0.0212766 0.00557932 0.25 0.0658436 2.718
3  της καρδιαγγειακής νόσου ||| cardiovascular disease ||| () (0) (1) |||
   (1) (2) ||| 0.0425532 0.00543132 0.666667 0.381907 2.718
4  την καρδιαγγειακή νόσο ||| cardiovascular disease ||| () (0) (1) ||| (1)
   (2) ||| 0.0851064 0.00523159 0.5 0.564943 2.718
5  οι ασθενείς για την καρδιαγγειακή νόσο ||| cardiovascular disease ||| ()
   () () () (0) (1) ||| (4) (5) ||| 0.0212766 7.79514e-10 0.25 0.564943 2.718
6  κίνδυνος καρδιαγγειακού ||| cardiovascular disease ||| () (0,1) ||| (1)
   (1) ||| 0.0212766 6.48728e-05 1 0.0658436 2.718
7  καρδιαγγειακών νοσημάτων . ||| cardiovascular disease ||| (0) (1) () |||
   (0) (1) ||| 0.0212766 2.59502e-06 1 0.347826 2.718
8  καρδιαγγειακών νοσημάτων ||| cardiovascular disease ||| (0) (1) ||| (0)
   (1) ||| 0.0212766 0.00104905 0.5 0.347826 2.718
9  καρδιαγγειακού κινδύνου ||| cardiovascular disease ||| (0,1) () ||| (0)
   (0) ||| 0.0212766 0.000194629 0.0357143 0.0658436 2.718
10 καρδιαγγειακού ||| cardiovascular disease ||| (0,1) ||| (0) (0) |||
   0.0425532 0.104904 0.08 0.0658436 2.718
```

Υπολογισμός Βέλτιστων Βαρών (Tuning)

```
47  # language model weights
48  [weight-l]
49  0.158323
50
51  # translation model weights
52  [weight-t]
53  0.018290
54  0.221347
55  0.124213
56  0.019937
57  -0.058909
```

Αποκωδικοποίηση (Decoding)

```
1  council regulation ( eec ) no 3812 / 85
2  of 20 december 1985
3  for the adaptation of regulations to milk and milk products the
   commission of the european communities ,
4  the commission of the european communities ,
5  having regard to the treaty establishing the european economic community
   ,
6  having regard to the proposal from the commission ( 1 ) , and in
   particular article 396 thereof ,
7  whereas pursuant to article 2 ( 3 ) of the directive , the uniform the
   measures referred to in article 396 of the act ; whereas these into force
   subject to and on the date of entry into force of the treaty , to milk
   and milk products , the following regulations :
8  - ( eec ) no 1098 / 68 of 27 july meet the application of the export to
   milk and milk products ( 1 ) , as last amended by regulation ( eec ) no
   2283 / 81 ( 3 ) ,
9  - ( eec ) no 1282 the commission of the provisions on the sale , at the
   entry of march , method and vegetables ( 1 ) , as last amended by
   regulation ( eec ) no 3474 / 80 ( 5 ) ,
```

Αξιολόγηση

```
1  CLUSTER_0_TEST: 26.51 (0.995) BLEU
2  TEST_1: 19.85 (0.883) BLEU
3  TEST_2: 19.57 (0.855) BLEU
4  TEST_3: 23.11 (0.891) BLEU
5  avg: 22.26 BLEU
```

MOSES

- Υλοποίηση μοντέλου στηριγμένου σε φράσεις
- Δυνατότητα προσθήκης επιπλέον πληροφορίας (κυρίως γλωσσολογικής)
- Αλγόριθμος αποκωδικοποίησης: Beam search
- Λογισμικό ανοιχτού κώδικα (sourceforge.net)

Προσαρμογή Γνωστικού Αντικειμένου (1)

- **Αμφισημία**
  1. “Note” («νότα» ή «σημείωση»;)
  2. “Log” («κορμός δένδρου» ή «ημερολόγιο»;)
- Τα **συμφραζόμενα** βοηθούν:
  1. “Play that **note** on the piano” v.

     “Can I borrow your organic chemistry **notes**?”
  2. “The Admiral needs the ship’s **log** book” v.

     “These **logs** will be sent to the paper mill”

Προσαρμογή Γνωστικού Αντικειμένου (2)

- Τρέχοντα συστήματα ΣΜΜ:
  - Λαμβάνουν υπόψη λίγες περιρρέουσες λέξεις
- Γνωρίζοντας το γνωστικό αντικείμενο του κειμένου προς μετάφραση, η άρση της αμφισημίας γίνεται ευκολότερη.
- Στόχος: Προσαρμογή του συστήματος ΣΜΜ σε διαφορετικά γνωστικά αντικείμενα.
- Αναμένεται βελτίωση και του ύφους.

Σώματα εκπαίδευσης

- 3 διαφορετικά σώματα εκπαίδευσης
  - Europarl
  - JRC-Acquis
  - Greek Paediatrics Journal

Europarl

	Characters	Words	Paragraphs
English	89,140,279	15,031,867	536,318
Greek	99,053,585	14,954,294	

JRC-Acquis

	Characters	Words	Paragraphs
English	41,917,357	6,465,374	227,007
Greek	46,013,152	6,649,694	

Greek Paediatrics Journal

	Characters	Words	Paragraphs
English	1,541,349	228,342	2,378
Greek	1,727,386	241,328	
Greek LM	21,778,272	3,109,470	146,242

Προσεγγίσεις

- Επιβλεπόμενη μάθηση (Μέθοδος 1)
  - Προϋπάρχοντα σώματα εκπαίδευσης
    - Ιατρικό
    - Europarl
- Μη επιβλεπόμενη μάθηση (Μέθοδος 2)
  - Δημιουργία συστάδων (clusters) από το JRC-Acquis
  - Cluto (tree clusterer)
  - 21 συστάδες (EUROVOC)

Αποτελέσματα Μεθόδου 1

E = Europarl M = Medical	TM (Translation Model)		LM (Language Model)		Average	Test file	
	E	M	E	M		TEST_MED_1	TEST_MED_2
Europarl + Med	Yes	Yes	Yes	Yes	**27.61**	34.77	20.45
Europarl + Med	Yes	Yes	No	Yes	**26.65**	33.98	19.32
Med (baseline)	No	Yes	No	Yes	**20.22**	27.14	13.31

Αποτελέσματα Μεθόδου 2

C18 = cluster_18 A = JRC-Acquis	TM (Translation Model)		LM (Language Model)		Test files			
	A	C18	A	C18	cluster_18_test	TEST_1	TEST_2	TEST_3
cluster_18 + JRC-Acquis	Yes	No	Yes	Yes	**31.05**	28.07	27.29	29.12
cluster_18 + JRC-Acquis	Yes	Yes	No	Yes	**30.10**	25.14	19.98	24.06
cluster_18 + JRC-Acquis	Yes	No	No	Yes	**30.18**	24.73	21.57	24.72
cluster_18 (baseline)	No	Yes	No	Yes	**23.29**	21.60	16.55	20.01

Παράδειγμα μεθόδου 1

- Baseline:
  - «given all spontaneous the existing data , appears that the microvascular complications may προληφθούν with επίτευξη ευγλυκαιμίας . this the κύριος and αυστηρός θεραπευτικός both the be επιτεύξιμος only ρυθμίζονται also when and the μεταγευματικές γλυκαιμικές excursions . »
- Μέθοδος 1:
  - «taking all existing data , microvascular complications can prevent achieving ευγλυκαιμίας . the main and strict therapeutic goal may be επιτεύξιμος only when adjusted also the μεταγευματικές γλυκαιμικές falls.»
- Μετάφραση αναφοράς:
  - «Taken together, the existing evidence suggests that microvascular complications can be prevented by sustained normoglycaemia. This main and stringent target of therapy may be achievable only when postprandial glycaemic excursions are also addressed. »

Συμπεράσματα

- Η εκπαίδευση ενός ΣΜΜ με ένα μικρό εξειδικευμένο σώμα σε συνδυασμό με ένα μεγάλο γενικό βελτιώνει σημαντικά την ποιότητα της μετάφρασης σε κείμενα του εξειδικευμένου σώματος.
- Ο υπολογισμός βέλτιστων βαρών (tuning) είναι ο πιο σημαντικός παράγοντας στη διαδικασία εκπαίδευσης
- Πρώτη χρήση του JRC-Acquis στο σύστημα ΣΜΜ «MOSES»

Βελτιώσεις

- Καλύτερος αλγόριθμος και διαδικασία συσταδοποίησης (clustering)
- Χρήση μηχανικής μάθησης για την αναγνώριση των προς μετάφραση κειμένων
- Χρήση λεξικού σε εξειδικευμένα κείμενα
- Καλύτερη ευθυγράμμιση των σωμάτων εκπαίδευσης

Ερωτήσεις

?