ΓΛΩΣΣΙΚΗ ΤΕΧΝΟΛΟΓΙΑ

ΑΥΤΟΜΑΤΗ (ΜΗΧΑΝΙΚΗ) ΜΕΤΑΦΡΑΣΗ

MACHINE TRANSLATION

Χρηματοδότηση

Το παρόν εκπαιδευτικό υλικό έχει αναπτυχθεί στα πλαίσια του εκπαιδευτικού έργου του διδάσκοντα.

Το έργο «**Ανοικτά Ακαδημαϊκά Μαθήματα στο Ιόνιο Πανεπιστήμιο**» έχει χρηματοδοτήσει μόνο τη αναδιαμόρφωση του εκπαιδευτικού υλικού.

Το έργο υλοποιείται στο πλαίσιο του Επιχειρησιακού Προγράμματος «Εκπαίδευση και Δια Βίου Μάθηση» και συγχρηματοδοτείται από την Ευρωπαϊκή Ένωση (Ευρωπαϊκό Κοινωνικό Ταμείο) και από εθνικούς πόρους.

Ευρωπαϊκή Ένωση
Ευρωπαϊκό Κοινωνικό Ταμείο
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΕΚΠΑΙΔΕΥΣΗ ΚΑΙ ΔΙΑ ΒΙΟΥ ΜΑΘΗΣΗ
επένδυση στην κοινωνία της γνώσης
ΥΠΟΥΡΓΕΙΟ ΠΑΙΔΕΙΑΣ & ΘΡΗΣΚΕΥΜΑΤΩΝ, ΠΟΛΙΤΙΣΜΟΥ & ΑΘΛΗΤΙΣΜΟΥ
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ
ΕΣΠΑ 2007-2013
πρόγραμμα για την ανάπτυξη
ΕΥΡΩΠΑΪΚΟ ΚΟΙΝΩΝΙΚΟ ΤΑΜΕΙΟ
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Άδειες Χρήσης

- Το παρόν εκπαιδευτικό υλικό υπόκειται σε άδειες χρήσης Creative Commons

Παραδοσιακή Μετάφραση

- Μετατροπή ενός κειμένου από μία φυσική γλώσσα σε μία άλλη
- Διαδικασία:
  - Διάβασμα του κειμένου σε μία φυσική γλώσσα
  - Κατανόηση του κειμένου
  - Παραγωγή κειμένου σε άλλη φυσική γλώσσα
- Δεν είναι μία απλή διαδικασία
  - Χρησιμοποιούνται άνθρωποι-ειδικοί για μεταφράσεις καλής ποιότητας

Μηχανική Μετάφραση (Machine Translation)

- Από τις πρώτες εφαρμογές της υπολογιστικής γλωσσολογίας (1950)
- Τεράστιες εμπορικές εφαρμογές
  - Η ΕΕ ξοδεύει πάνω από 1 δις € σε κόστη μετάφρασης κάθε χρόνο
- Πολύ δύσκολο πρόβλημα ειδικά για μετάφραση:
  - εντελώς αυτοματοποιημένη
  - πραγματικού χρόνου
  - ανοιχτού λεξιλογίου

Εκτός από εμπορικό, και ερευνητικό ενδιαφέρον

- Συνδυάζει πολλές τεχνολογίες επεξεργασίας φυσικής γλώσσας:
  - Αναγνώριση μερών του λόγου
  - Συντακτική ανάλυση
  - Σύνθεση
  - Άρση αμφισημίας λέξης
  - Αναγνώριση ονομάτων –οντοτήτων
  - Επίλυση ασάφειας αναφορών
  - Κατανόηση φυσικής γλώσσας
  - Αναπαράσταση γνώσης του κόσμου

Ιστορική Αναδρομή

- 1950's
  - Έντονη δραστηριότητα - Απλά συστήματα - Ανταγωνισμός Αμερικής-Ρωσίας
- 1966: ALPAC report
  - Αρνητική αναφορά για την πρόοδο της έρευνας
  - Περικοπή κονδυλίων
- 1966-1975
  - Συστήματα 2ης γενιάς: πιο περίπλοκα από γλωσσολογικής και υπολογιστικής άποψης
- 1975-1985
  - Οι πρώτες επιτυχίες: Météo - Systran
- 1985-σήμερα
  - Eurotra, Ιαπωνικά συστήματα, Επανέναρξη έρευνας στην Αμερική
  - Πρώτα εμπορικά συστήματα (για PCs)
  - Στατιστικές μέθοδοι
  - Αμφιλεγόμενα αποτελέσματα, κριτική

Σήμερα

- Υπάρχουν αρκετά αξιόπιστα εμπορικά συστήματα
  - Μερικά αρκετά φτηνά ($50)
  - Χωρίς ιδιαίτερες υπολογιστικές απαιτήσεις (PC)
- Ελεύθερη μετάφραση μέσω WWW
  - Μετάφραση ιστοσελίδων και email
  - Η χαμηλή ποιότητα μετάφρασης είναι αποδεκτή
  - Καλύπτεται ένα μικρό μέρος των φυσικών γλωσσών
- Έρευνα στη μετάφραση προφορικού λόγου
  - Verbmobil

Χρησιμότητα Μηχανικής Μετάφρασης

- Σε εργασίες όπου μία πρόχειρη μετάφραση είναι επαρκής
  - Μετάφραση ιστοσελίδων
  - Διαγλωσσική ανάκτηση πληροφορίας
- Σε εργασίες όπου μπορεί να γίνει διόρθωση της αυτόματης μετάφρασης από κάποιον άνθρωπο-ειδικό
  - Human-assisted machine translation
- Σε εργασίες όπου επεξεργάζονται υπογλώσσες
  - Δελτία καιρού
  - Εγχειρίδια συσκευών

Babel Fish (http://babelfish.altavista.com/)

Babel Fish: Greek → English

- Καλώς ήρθατε στο Τμήμα Πληροφορικής του Ιονίου Πανεπιστημίου. Το Τμήμα Πληροφορικής δημιουργήθηκε στο πλαίσιο του ΕΠΕΑΕΚ και λειτουργεί από το ακαδημαϊκό έτος 2004-05. Το Τμήμα δέχεται φοιτητές/τριες από το 2ο και 4ο επιστημονικό πεδίο και έχει ως αντικείμενο τόσο τη θεωρητική όσο και την εφαρμοσμένη Πληροφορική.

- Well you came in the Department of Information technology of Ionian University. The Department of Information technology was created in the frame of SPECIAL TRAINING PROGRAM and functions from the academic year 2004-05. the Department accepts students/trjes from the 2nd and 4th scientific field and has as object so much the theoretical what applied Information technology.

Babel Fish: English → Greek

- Welcome to the Department of Informatics of the Ionian University. The Department of Informatics was founded by the Ministry of National Education and Religious Affairs in 2004 and its scope covers Theoretical as well as Applied Informatics.

- Υποδοχή στο τμήμα πληροφορικής του ιόνιου πανεπιστημίου. Το τμήμα πληροφορικής ιδρύθηκε από το Υπουργείο εθνικής παιδείας και θρησκευτικών υποθέσεων το 2004 και το πεδίο του καλύπτει τη θεωρητική καθώς επίσης και εφαρμοσμένη πληροφορική.

Προκλήσεις στην Αυτόματη Μετάφραση (1/3)

Οι φυσικές γλώσσες διαφέρουν σε πολλά μεταξύ τους

- Μορφολογικές διαφορές
  - the → ο, η, το, τα, του, της, των, ...
- Αντωνυμίες
  - Σε πολλές γλώσσες (μορφολογικά πλούσιες) η αντωνυμία-υποκείμενο στην πρόταση εννοείται και τα μορφολογικά της χαρακτηριστικά καθορίζονται από την μορφολογία του ρήματος
  - Η κατάληξη του ρήματος στα Ισπανικά δείχνει ποιά αντωνυμία εννοείται
    - -o = I
    - -as = you
    - -a = he/she/it !!! (Ποιο θα επιλεγεί;)
    - -amos = we
    - -an = they

Προκλήσεις στην Αυτόματη Μετάφραση (2/3)

- Συντακτικές διαφορές (διάταξη των όρων)
  - language use → χρήση γλώσσας
    english(N1 N2) → greek(N2 N1)
  - the new house → la casa nueva
    english(DT J N) → spanish(DT N J)
  - IBM bought Lotus → IBM Lotus bought
    english(SUBJ V OBJ) → japanese(SUBJ OBJ V)
- Διαφορές στην έννοια των λέξεων
  - wall → Wand (inside), Mauer (outside)
- Διαφορές στην έκφραση
  - Αγγλικά: *I am hungry* (είμαι πεινασμένος)
  - Γερμανικά: *Ich habe Hunger* (έχω πείνα)
  - Ελληνικά: *Πεινάω*

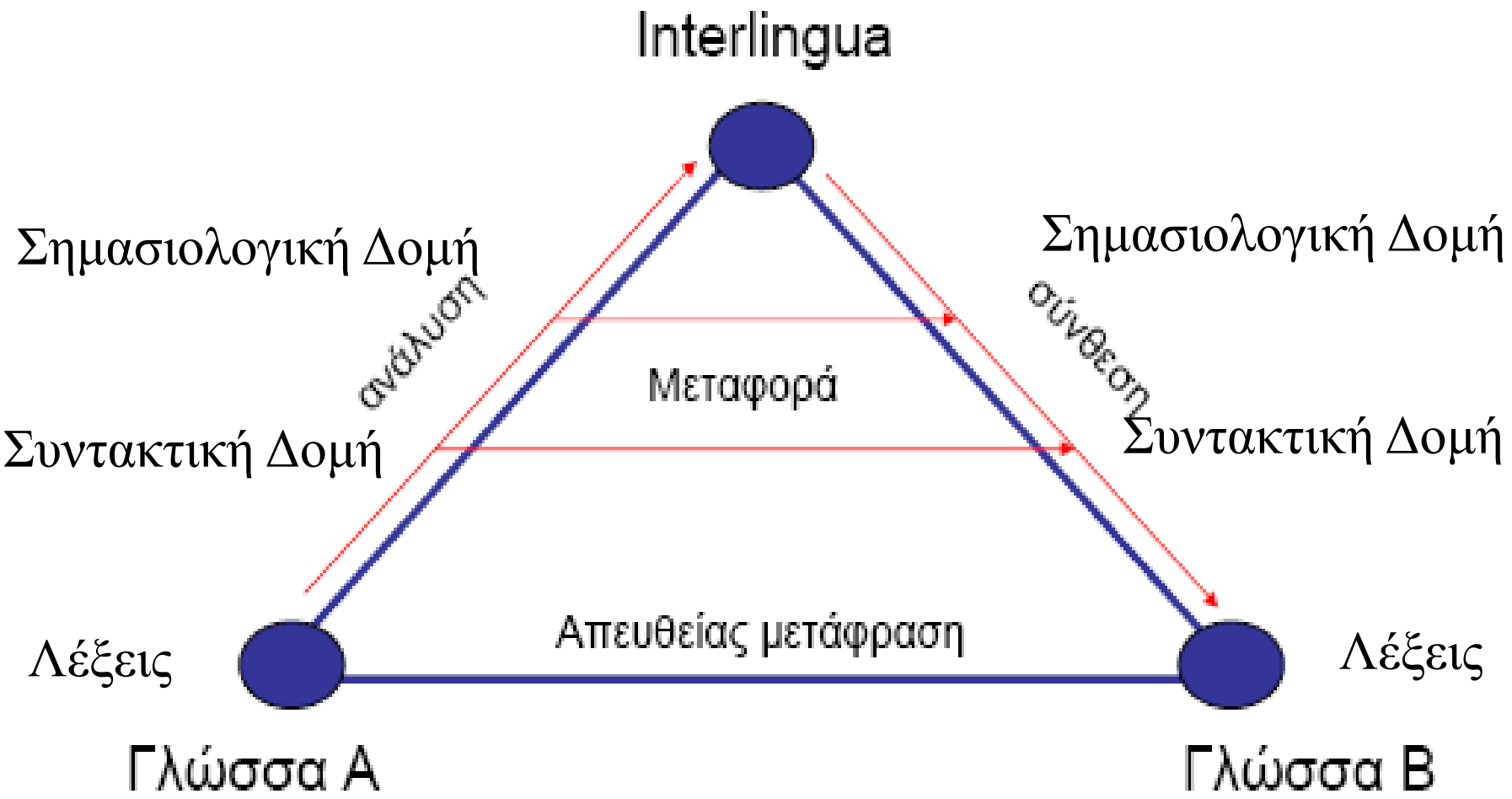
Προκλήσεις στην Αυτόματη Μετάφραση (3/3)

- Ο χρόνος των ρημάτων
  - I have been playing the piano for three years
  - Παίζω πιάνο τρία χρόνια
- Ιδιωματισμοί
  - He kicked the bucket → Πέθανε
  - She has always been a lame duck → Πάντα ήταν άχρηστη/βαρετή/ανίκανη

Κλασσικά Μοντέλα Μηχανικής Μετάφρασης

- Interlingua
  - Για τη μετάφραση από μία γλώσσα Α σε μία γλώσσα Β χρησιμοποιείται ως ενδιάμεσο μία ουδέτερη γλώσσα (interlingua - αναπαράσταση νοήματος)
- Transfer (μεταφορά)
  - Για τη μετάφραση από μία γλώσσα Α σε μία γλώσσα Β ορίζεται μία διαδικασία ανάλυσης, μεταφοράς και σύνθεσης
- Direct (word-for-word) translation (απευθείας μετάφραση)
  - Για τη μετάφραση από μία γλώσσα Α σε μία γλώσσα Β γίνεται απευθείας μεταφορά από την μία στην άλλη

Τρίγωνο Vauquois



Μοντέλο Interlingua (1/2)

- Interlingua: γλώσσα αναπαράστασης του νοήματος μιας πρότασης
  - Ο Γιάννης πρέπει να μην πάει → OBLIGATORY(NOT(GO(JOHN)))
  - Ο Γιάννης μπορεί να μην πάει → NOT(OBLIGATORY(GO(JOHN)))
- Με αυτή την αναπαράσταση παράγεται η πρόταση σε άλλη γλώσσα
- Πλεονέκτημα: Μπορεί να γίνει μετάφραση μεταξύ οποιωνδήποτε γλωσσών και χρειάζεται μόνο η υλοποίηση εργαλείων ανάλυσης/σύνθεσης ξεχωριστά για κάθε γλώσσα

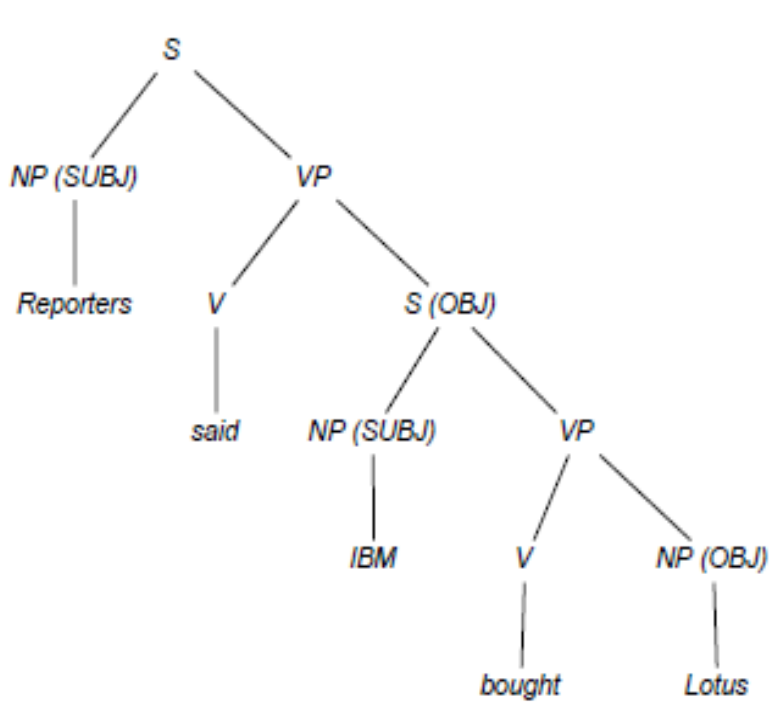
Μοντέλο Interlingua (2/2)

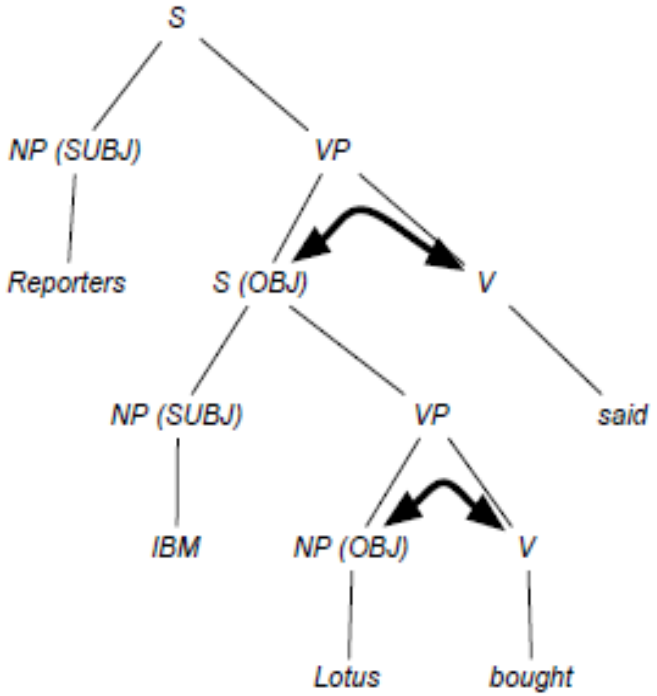
- Μειονεκτήματα:
  - Απαιτεί πολύ προσεκτικό σχεδιασμό της interlingua
  - Απαιτεί για κάθε γλώσσα την δυνατότητα μετάβασης από και προς την interlingua
  - Το μεγαλύτερο πρόβλημα είναι ότι πρέπει να αποσαφηνίσουμε εντελώς το νόημα σε κάθε περίπτωση
- Η interlingua μπορεί να είναι είτε μία τεχνητή γλώσσα (αναπαράσταση νοήματος) είτε μία τρίτη φυσική γλώσσα

Μοντέλο Μεταφοράς (1/2)

- Πραγματοποιείται ανάλυση του κειμένου-εισόδου
- Εφαρμόζονται κανόνες μετασχηματισμού της γλωσσολογικής δομής της εισόδου στην γλωσσολογική δομή της εξόδου
- Από την συντακτική δομή της εξόδου, παράγεται η πρόταση εξόδου (σύνθεση)
- Η διαδικασία της σύνθεσης μπορεί να παράγει πολλές εναλλακτικές εξόδους από τις οποίες επιλέγεται η καλύτερη
- Πλεονέκτημα: Αντιμετωπίζει το πρόβλημα της διάταξης
- Μειονέκτημα: Πρέπει να κατασκευαστούν κανόνες συντακτικών μετασχηματισμών για κάθε ζεύγος γλωσσών

Μοντέλο Μεταφοράς (2/2)





- Η πρόταση-πηγή αναλύεται συντακτικά
- Πραγματοποιούνται οι απαραίτητες αναδιατάξεις των όρων
- Μεταφράζονται οι λέξεις

Απευθείας Μετάφραση

- Πλεονεκτήματα
  - Συμπεριλαμβάνει μόνο μορφολογική ανάλυση
  - Γίνεται απλή μεταφορά των λέξεων από τη μία γλώσσα στην άλλη με χρήση μεγάλου δίγλωσσου λεξικού
- Μειονεκτήματα
  - Η πρόταση-μετάφραση απαιτεί αναδιάταξη των όρων
    - Σειρά ουσιαστικών-επιθέτων
    - Πρόσθεση/αφαίρεση άρθρων, προθέσεων
    - Μορφολογική σύνθεση

Συνδυασμός Μοντέλων

- Το Systran αποτελείται από 3 συστατικά:
  - Ανάλυση
    - Μορφολογική ανάλυση και αναγνώριση ΜΤΛ
    - Ανίχνευση ονοματικών και προθετικών φράσεων
    - Ρηχή συντακτική ανάλυση
  - Μεταφορά
    - Μετάφραση ιδιωματισμών
    - Άρση Αμφισημίας Λέξεων
    - Προσάρτηση προθετικών φράσεων
  - Σύνθεση
    - Χρήση δίγλωσσου λεξικού για την μετάφραση των λέξεων
    - Αναδιάταξη λέξεων
    - Μορφολογική σύνθεση

Στοχαστικές Μέθοδοι

- Σε αντίθεση με τις ορθολογιστικές προσεγγίσεις βάσει κανόνων, οι στοχαστικές μέθοδοι μετάφρασης στηρίζονται στα δεδομένα (κείμενα)
- Είναι η πιο ελπιδοφόρα προσέγγιση αφού υπάρχει πλέον
  - Μεγάλη αποθηκευτική ικανότητα
  - Μεγάλη επεξεργαστική ισχύς
  - Τεράστιες ποσότητες διαθέσιμων δεδομένων
- Δύο βασικές προσεγγίσεις
  - Στατιστική μηχανική μετάφραση
  - Μετάφραση βάσει παραδειγμάτων

Στατιστική Μηχανική Μετάφραση

- Ευθυγράμμισε αυτόματα λέξεις (word-based) ή/και φράσεις (phrase-based) στις προτάσεις ενός παράλληλου σώματος κειμένων
- Υπολόγισε τις πιθανότητες μετάφρασης εκπαιδεύοντας ένα στατιστικό μοντέλο με το παράλληλο σώμα κειμένων
- Βρες την πιο πιθανή πρόταση στην γλώσσα B (γλώσσα-στόχος), δεδομένης μιας πρότασης στην γλώσσα A (γλώσσα-πηγή)
  - Την πρόταση B που μεγιστοποιεί την p(B|A)

Στατιστική Μηχανική Μετάφραση

- Πλεονεκτήματα
  - Αντιμετωπίζει την αμφισημία
  - Αντιμετωπίζει ιδιωματισμούς
  - Απαιτεί ελάχιστη ανθρώπινη παρέμβαση
    - Υλοποίηση χωρίς μεγάλο οικονομικό και χρονικό κόστος
  - Μπορεί να υλοποιηθεί για οποιοδήποτε ζεύγος γλωσσών που διαθέτει δεδομένα εκπαίδευσης
- Μειονεκτήματα
  - Δεν αντιμετωπίζει ρητά συντακτικές δομές

Παράλληλο Σώμα Κειμένων (Parallel Corpus)

what is more , the relevant cost dynamic is completely under control.	im übrigen ist die diesbezügliche kostenentwicklung völlig unter kontrolle .
sooner or later we will have to be sufficiently progressive in terms of own resources as a basis for this fair tax system .	früher oder später müssen wir die notwendige progressivität der eigenmittel als grundlage dieses gerechten steuersystems zur sprache bringen .
we plan to submit the first accession partnership in the autumn of this year .	wir planen , die erste beitrittspartnerschaft im herbst dieses jahres vorzulegen .
it is a question of equality and solidarity .	hier geht es um gleichberechtigung und solidarität .
the recommendation for the year 1999 has been formulated at a time of favourable developments and optimistic prospects for the european economy .	die empfehlung für das jahr 1999 wurde vor dem hintergrund günstiger entwicklungen und einer für den kurs der europäischen wirtschaft positiven perspektive abgegeben .
that does not , however , detract from the deep appreciation which we have for this report .	im übrigen tut das unserer hohen wertschätzung für den vorliegenden bericht keinen abbruch .

Ευθυγράμμιση - Alignment

- Παράλληλα κείμενα
  - Τα ίδια κείμενα γραμμένα στις δύο γλώσσες
- Επιπλέον, τα κείμενα πρέπει να είναι ευθυγραμμισμένα (aligned)
  - Σε ποια πρόταση (ή προτάσεις) μιας γλώσσας αντιστοιχεί μια πρόταση της άλλης γλώσσας
  - Σε ποια λέξη/φράση μιας γλώσσας αντιστοιχεί μια λέξη/φράση της άλλης γλώσσας

Παραδείγματα Παράλληλων Κειμένων

- Πρακτικά Καναδικής Βουλής
- Επίσημη Εφημερίδα Ευρωπαϊκής Ένωσης
- Αναφορές Ηνωμένων Εθνών
- Εγχειρίδια χρήσης συσκευών
- Νομοθεσία Hong-Kong, Macao
- ...

Πιθανότητες

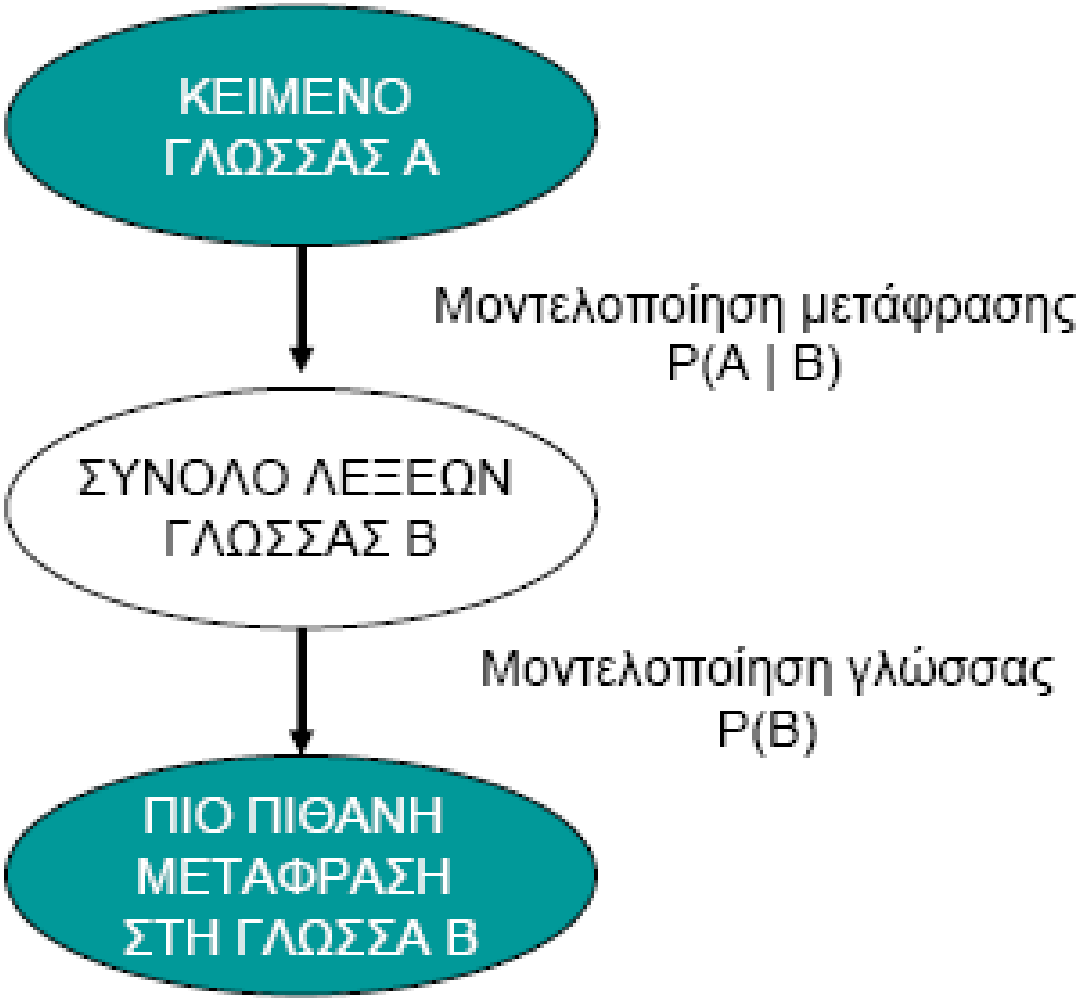
- Για μια πρόταση στην γλώσσα A, θέλουμε την πρόταση στην γλώσσα B που μεγιστοποιεί την πιθανότητα p(B|A)
- P(B|A) =P(B) * P(A|B) / P(A)
- B= argmax_B P(B)*P(A|B)
- P(B): μοντέλο γλώσσας
- P(A|B): μοντέλο μετάφρασης

Το P(A) είναι σταθερό μια και ψάχνω την καλύτερη πρόταση της γλώσσας B για μια σταθερή πρόταση της γλώσσας A

Μοντέλο γλώσσας (Language Model) Μοντέλο μετάφρασης (Translation model)

- Μοντέλο γλώσσας
  - Αποδίδει μεγαλύτερες πιθανότητες σε γραμματικά/συντακτικά σωστές προτάσεις
  - Οι πιθανότητες αυτές υπολογίζονται με μονόγλωσσα σώματα κειμένων
- Μοντέλο Μετάφρασης
  - Αποδίδει μεγαλύτερες πιθανότητες σε προτάσεις που έχουν παρόμοιο νόημα
  - Οι πιθανότητες υπολογίζονται με χρήση δίγλωσσων σωμάτων κειμένων

Στατιστική Μηχανική Μετάφραση



Παράδειγμα: Γαλλικά → Αγγλικά

On voit Jon à la télévision

	good English? P(E)	good match to French? P(F|E)
Jon appeared in TV.		✓
Appeared on Jon TV.		✓
In Jon appeared TV.		
Jon is happy today.	✓	
Jon appeared on TV.	✓	✓
TV appeared on Jon.	✓	
TV in Jon appeared.		
Jon was not happy.	✓	

Μοντέλο Γλώσσας

- Προσπαθεί να εξασφαλίσει ότι οι λέξεις στην πρόταση-στόχο θα εμφανιστούν με την σωστή διάταξη
- Δεν είναι πρακτικό να υπολογίσουμε τη πιθανότητα όλων των δυνατών συνδυασμών λέξεων
- Συνήθως υπολογίζεται με μοντέλο τριγράμμου

p(I like bungee jumping off high bridges) =
p(I | <s> <s>) *
p(like | I <s>) *
p(bungee | I like) *
p(jumping | like bungee) *
p(off | bungee jumping) *
p(high | jumping off) *
p(bridges | off high) *
p(</s> | high bridges) *
p(</s> | bridges </s>)

Unigram probabilities

$$p(w_1) = \frac{count(w_1)}{total\ words\ observed}$$

Bigram probabilities

$$p(w_2|w_1) = \frac{count(w_1 w_2)}{count(w_1)}$$

Trigram probabilities

$$p(w_3|w_1 w_2) = \frac{count(w_1 w_2 w_3)}{count(w_1 w_2)}$$

Δίγραμμα
Μεγάλη αύξηση του N

- Μοντέλο bigrams: κάθε λέξη εξαρτάται μόνο από την προηγούμενή της
  - $P(w_1,w_2,\ldots,w_n) = \Pi\, P(w_i \mid w_{i-1})$
  - Πολλές φορές δεν αρκεί αυτό το μοντέλο:

    π.χ. “I hire the men who is good pilots”
- Αν αυξήσουμε το *n* πολύ, υπάρχει ο κίνδυνος μερικοί συνδυασμοί να έχουν πιθανότητα 0

Υπολογισμός πιθανοτήτων

- Όσο μεγαλύτερες ακολουθίες λέξεων χρησιμοποιώ για τον υπολογισμό των πιθανοτήτων, τόσο πιο απίθανο είναι να συναντήσω αυτές τις ακολουθίες στα δεδομένα
- Πρόβλημα σπάνιων δεδομένων (sparse data)
- Λύση: backing off (smoothing – εξομάλυνση)
- Χρησιμοποιώ συνδυασμό unigrams + bigrams + trigrams με αντίστοιχο βάρος στο καθένα

$$
\begin{aligned}
&.8 * p(w_3|w_1 w_2) + \\
&.15 * p(w_3|w_2) + \\
&.049 * p(w_3) + \\
&.001
\end{aligned}
$$

Μοντέλο Μετάφρασης

- Δεδομένης μιας μετάφρασης (πρόταση B), ποια η πιθανότητα να προέρχεται από την πρόταση A στην γλώσσα-πηγή;
- P(A|B) = count (A,B) / count(B)
- Αδύνατο, γιατί αποκλείεται να έχω αρκετά δεδομένα ώστε να έχω μετρήσεις για ολόκληρες προτάσεις
- Για αυτό σπάω τις προτάσεις σε υπο-συστατικά
  - Λέξεις (word-based SMT)
  - Φράσεις (phrase-based SMT)

Ευθυγράμμιση (Alignment)

- Πρέπει για το κάθε υπο-συστατικό (λέξη ή φράση) να βρεθεί η μετάφρασή του ανάμεσα στα ζευγάρια προτάσεων του παράλληλου σώματος
- Δεδομένης της λέξης *worked*

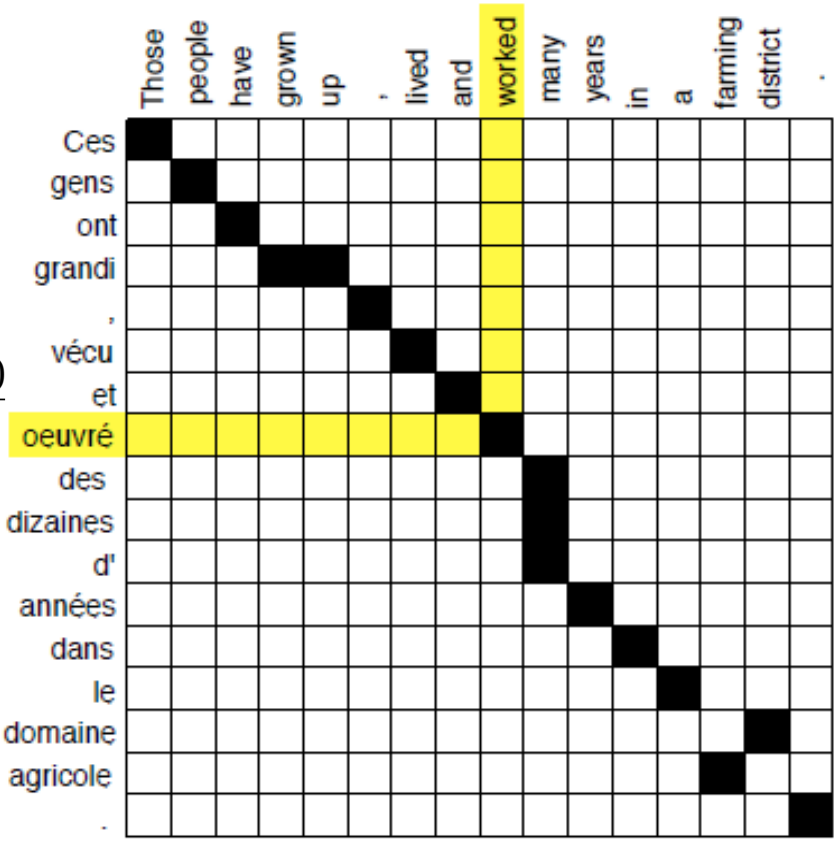
ποια η πιθανότητα να έχει προκύψει από μετάφραση της λέξης

travaille? $t(\text{travaille} \mid \text{worked}) = \frac{\text{count}(\text{travaille}, \text{worked})}{\text{count}(\text{worked})}$

marche? $t(\text{marche} \mid \text{worked}) = \frac{\text{count}(\text{marche}, \text{worked})}{\text{count}(\text{worked})}$

oeuvre? …

fonctionne? …



Πχ αν από τις 100 φορές που εμφανίζεται συνολικά το *worked* στο παράλληλο σώμα, τις 13 είναι μετάφραση του oeuvre, τότε t(oeuvre |worked)=0.13

Πιθανότητες Ευθυγράμμισης

$$p(a, A \mid B) = \prod_{j=1}^{m} t(A_j \mid B_i)$$

πιθανότητες ευθυγράμμισης
alignment probabilities

Η j λέξη στην πρόταση της γλώσσας-πηγή

Η λέξη στη γλώσσα στόχο που έχει προκύψει από την ευθυγράμμιση με την λέξη Aj

$$p(A \mid B) = \sum_{\alpha} p(a, A \mid B)$$

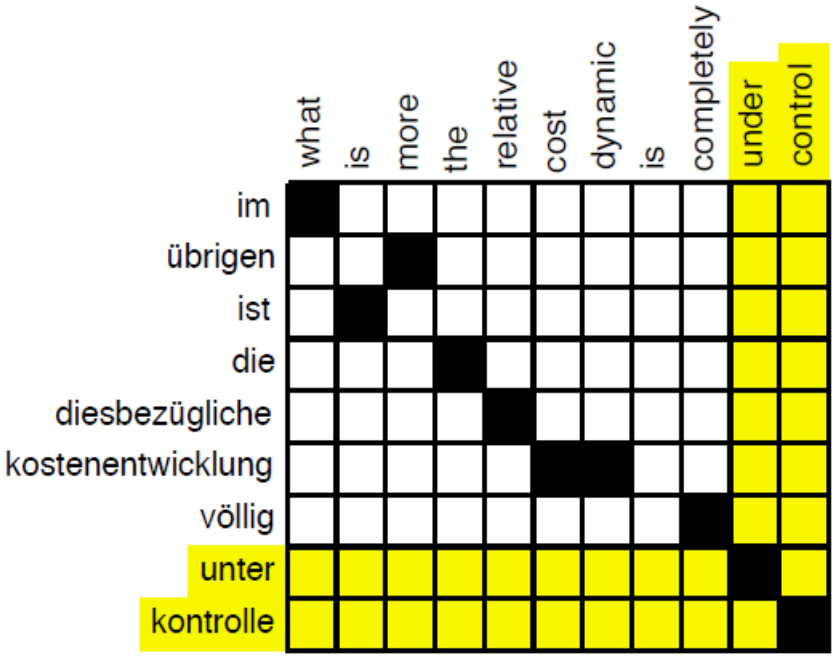
πιθανότητες μετάφρασης
translation probabilities

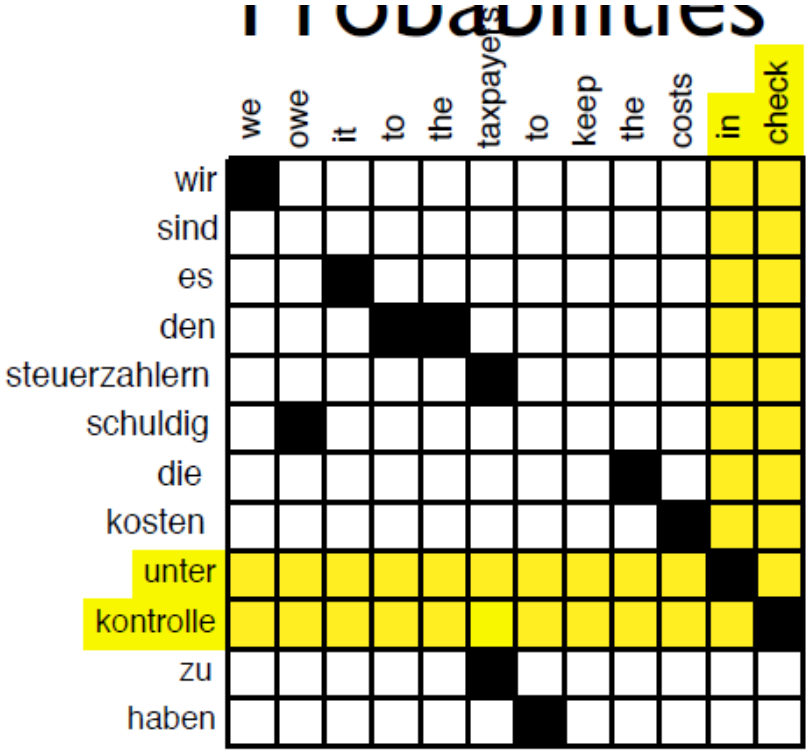
Υπάρχει περίπτωση να μπορεί να παραχθεί η ίδια πρόταση στην γλώσσα-στόχο από την ίδια πρόταση στην γλώσσα-πηγή με διαφορετικούς συνδυασμούς ευθυγραμμίσεων των λέξεων. **Οπότε η πιθανότητα της μετάφρασης Β να έχει προέλθει από την Α είναι το άθροισμα των πιθανοτήτων όλων των πιθανών ευθυγραμμίσεων.**

Ο υπολογισμός του t(Aj|Bi)

- Για να μπορέσουν να υπολογιστούν οι πιθανότητες μετάφρασης πρέπει τα δεδομένα μου να είναι ευθυγραμμισμένα λέξη προς λέξη.
- Δυστυχώς αυτό είναι πολύ σπάνιο

Phrase-based SMT





Πίνακες φράσεων

Εξαντλητικοί πίνακες με φράσεις στην γλώσσα-πηγή, που συνοδεύονται με τις πιθανές μεταφράσεις τους στην γλώσσα-στόχο, εμπλουτισμένες με πιθανότητες.

das thema	the issue	.51
	the point	.38
	the subject	.21

Ευθυγράμμιση Κειμένων (Text Alignment)

- Μία πρόταση στο κείμενο μιας γλώσσας δεν αντιστοιχεί πάντα σε μία πρόταση στο κείμενο μιας άλλης γλώσσας
  - Αν και αυτό συμβαίνει κατά 90%
- Οι διαφορετικές αντιστοιχήσεις καλούνται beads
  - 1:1, 1:0, 0:1, 2:1, 1:2, 2:2
- Η σειρά των προτάσεων μπορεί να αλλάζει κατά τη μετάφραση
- Ευθυγράμμιση προτάσεων - λέξεων

Προσεγγίσεις στην Ευθυγράμμιση Προτάσεων

- Βάσει μήκους προτάσεων
  - Οι μικρές προτάσεις αντιστοιχούν σε μικρές και οι μεγάλες σε μεγάλες
- Λεξιλογικές μέθοδοι
  - Η αντιστοίχηση των προτάσεων γίνεται βάσει λεξιλογικής πληροφορίας
- Ευθυγράμμιση offsets με τεχνικές επεξεργασίας σήματος
  - Ευθυγράμμιση συγκεκριμένων σημείων και όχι προτάσεων

Ευθυγράμμιση Λέξεων

- Γίνεται με βάση ένα παράλληλο κείμενο ευθυγραμμισμένο ως προς τις προτάσεις
- Εξαγωγή λεξιλογίου για την κάθε γλώσσα
- Για κάθε ζευγάρι λέξεων, υπολογισμός της πιθανότητας να αντιστοιχούν σε αντίστοιχους όρους
- Γονιμότητα λέξης (fertility): δεν είναι όλες οι αντιστοιχίσεις μία προς μία
  - Μερικές λέξεις έχουν πολλαπλές μεταφράσεις (*the* → *ο*, *η*, *το*, ...)
  - Μερικές λέξεις δεν έχουν καθόλου μετάφραση (*is running* → *τρέχει*, *is* → ∅)
  - Μερικές λέξεις μεταφράζονται με περισσότερες λέξεις (*απογειώνομαι* → *take off*)

Παράδειγμα (Αγγλικά → Γαλλικά)

Fertility:

The	(→ Les)	= 1
not	(→ ne .. pas)	= 2
be	(→ ∅)	= 0

Πιθανότητες για Ζεύγη Λέξεων (Από τα Πρακτικά της Καναδικής Βουλής)

English: the

French	P	fertility	P
le	.610	1	.871
la	.178	0	.124
l'	.083	2	.004
les	.023		
ce	.013		
il	.012		
de	.009		
a	.007		
que	.007		

English: not

French	P	fertility	P
pas	.469	2	.758
ne	.460	0	.133
non	.024	1	.106
faux	.006		
plus	.002		
ce	.002		
que	.002		
jamais	.002		

Ενσωμάτωση και του φαινομένου της γονιμότητας στον υπολογισμό των πιθανοτήτων του μοντέλου μετάφρασης

Μέχρι τώρα: $$p(a, A \mid B) = \prod_{j=1}^{m} t(A_j \mid B_i)$$

Πιθανότητες γονιμότητας: *n(1|'house')* – Η πιθανότητα η λέξη 'house' να ευθυγραμμίζεται με ακριβώς μια λέξη στην γλώσσα Α κάθε φορά που εμφανίζεται στα παράλληλα κείμενα εκπαίδευσης.

Οπότε: $$p(a, A \mid B) = \prod_{j=1}^{m} t(A_j \mid B_i) \prod_{i=1}^{l} n(f(B_i) \mid B_i)$$

f(Bi): η γονιμότητα της λέξης *Bi*

Παραμόρφωση (Distortion)

- Οι μεταφρασμένες λέξεις δεν εμφανίζονται με την ίδια σειρά
- Το μοντέλο μετάφρασης συμπεριλαμβάνει πιθανότητες «παραμόρφωσης»
  - P(2 | 5): Η πιθανότητα η λέξη w_A να εμφανιστεί στη θέση 2 όταν η w_B είναι στη θέση 5
  - P(2 | 5,4,6): Η πιθανότητα μία w_A στη θέση 2 να αντιστοιχεί με μία w_B στη θέση 5 όταν η πρόταση Α έχει 6 λέξεις και η πρόταση Β 4 λέξεις

Ενσωμάτωση και του φαινομένου της παραμόρφωσης στον υπολογισμό των πιθανοτήτων του μοντέλου μετάφρασης

Μέχρι τώρα: $$p(a, A \mid B) = \prod_{j=1}^{m} t(A_j \mid B_i) \prod_{i=1}^{l} n(f(B_i) \mid B_i)$$

Τώρα:

$$p(a, A \mid B) = \prod_{j=1}^{m} t(A_j \mid B_i) \prod_{i=1}^{l} n(f(B_i) \mid B_i) \prod_{j=1}^{m} d(j \mid a_j, l, m)$$

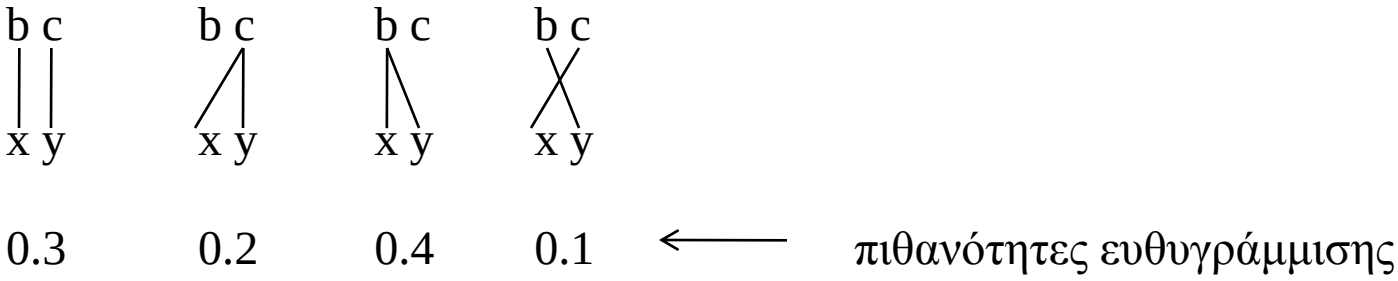
d(j|aj,l,m): η πιθανότητα η λέξη στην θέση *j* στην πρόταση-πηγή να μεταφραστεί σε μια λέξη που θα βρίσκεται στην θέση *αj* στην γλώσσα στόχο.

l: αριθμός λέξεων πρότασης στη γλώσσα-στόχο

m: αριθμός λέξεων πρότασης στη γλώσσα-πηγή

Το πρόβλημα: η κότα και το αυγό (1/2)

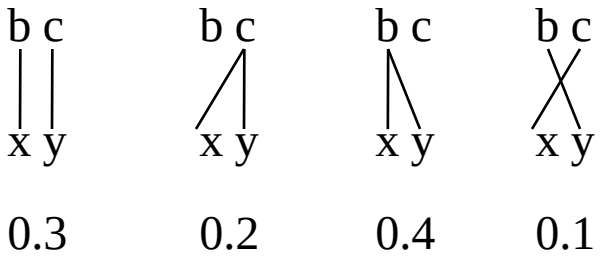
- Εάν έχω τις παραμέτρους του στατιστικού μοντέλου, μπορώ με βάση τον προηγούμενο τύπο να υπολογίσω πιθανότητες ευθυγράμμισης.
- Εάν έχω πιθανότητες ευθυγράμμισης, τότε μπορώ να υπολογίσω τις παραμέτρους του στατιστικού μοντέλου. Πχ



Η πιθανότητα η λέξη b να παράγει ακριβώς μια λέξη ως μετάφρασή της
n(1|b) = count(1|b)/(count(0|b)+count(1|b)+count(2|b)) = (0.3+0.1) / (0.2+0.4+0.4) = 0.4 **(fractional counts)**

Το πρόβλημα: η κότα και το αυγό (2/2)

- Αντίστοιχα μπορώ να υπολογίσω με **fractional counts** τις πιθανότητες μετάφρασης t(y|b)



t(y|b) = count(y|b) / (count(x|b)+count(-|b)+count(y|b)+count(xy|b))

= 0.1 / (0.3+0.2+0.1+0.4) = 0.1

Για να υπολογίσουμε το *p(A|B)*, δηλ. το *p(α, A|B)* χρειαζόμαστε τις στατιστικές παραμέτρους. Για να έχουμε τις παραμέτρους χρειαζόμαστε πιθανότητες ευθυγράμμισης, και για αυτές χρειαζόμαστε τις παραμέτρους → φαύλος κύκλος

Λύση: Ο Αλγόριθμος EM (Expectation Maximization)

- Δώσε uniform αρχικές τιμές στις παραμέτρους
  - Εάν πχ υπάρχουν 40,000 λέξεις στο λεξιλόγιο της γλώσσας Α, τότε t(A|B)=1/40.000 για κάθε ζεύγος λέξεων.
  - Επιλέγω μια τυχαία τιμή και για την γονιμότητα (πχ 0.15), κοινή για όλες τις λέξεις της γλώσσας Β
- Από τις παραμέτρους αυτές υπολόγισε πιθανότητες ευθυγράμμισης
- Από τις πιθανότητες ευθυγράμμισης υπολόγισε καινούριες τιμές στις παραμέτρους.
- Από τις καινούριες παραμέτρους υπολόγισε καινούριες πιθανότητες ευθυγράμμισης κλπ κλπ
- Μέχρι να επιτευχθεί σύγκλιση

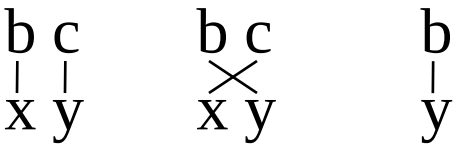
Παράδειγμα

- Έστω το σώμα κειμένων

b c – x y

b – y

Οι πιθανές ευθυγραμμίσεις είναι



(υποθέτω για όλες τις λέξεις γονιμότητα 1 και καθόλου παραμόρφωση)

$$p(a, A \mid B) = \prod_{j=1}^{m} t(A_j \mid B_i)$$

Βήμα 1. Δώσε στις παραμέτρους uniform βάρη

t(x|b)=0.5 t(x|c)=0.5

t(y|b)=0.5 t(y|c)=0.5

Παράδειγμα (συν)

Βήμα 2. Υπολόγισε το p(α,f|e) για κάθε ευθυγράμμιση

p(α, A|B)=0.5*0.5=0.25

p(α, A|B)=0.5*0.5=0.25

p(α, A|B)=0.5

Βήμα 3. Με fractional counts υπολόγισε πιθανότητες ευθυγράμμισης

p(α| A,B)=0.25/(0.25+0.25)=0.5

p(α| A,B)=0.25/(0.25+0.25)=0.5

p(α| A,B)=0.5/0.5=1

Παράδειγμα (συν)

Βήμα 4. Υπολόγισε με fractional counts ξανά τις παραμέτρους

t(x|b) = count(x|b) / (count(x|b)+count(y|b))= 0.5 / (0.5+0.5+1) = 0.25

t(y|b)=count(y|b)/(count(x|b)+count(y|b))=(0.5+1)/ (0.5+0.5+1)=0.75

t(x|c) = count(x|c) / (count(x|c)+count(y|c))= 0.5 / (0.5+0.5) = 0.5

t(y|c) = count(y|c) / (count(x|c)+count(y|c))= 0.5 / (0.5+0.5) = 0.5

Βήμα 5. Υπολόγισε καινούριες πιθανότητες ευθυγράμμισης

p(α, A|B)=0.25*0.5=0.125

p(α, A|B)=0.75

p(α, A|B)=0.75*0.5=0.375

Παράδειγμα (συν)

Βήμα 6. Με fractional counts υπολόγισε καινούριες πιθανότητες ευθυγράμμισης

p(α| A,B)=0.125/(0.125+0.375)=0.25

p(α| A,B)=0.375/(0.125+0.375)=0.75

p(α| A,B)=1

Βήμα 7. Υπολόγισε με fractional counts ξανά τις παραμέτρους

t(x|b) = count(x|b) / (count(x|b)+count(y|b))= 0.25 / (0.25+0.75+1) = 0.125
t(y|b)=count(y|b)/(count(x|b)+count(y|b))=(0.75+1)/ (0.25+0.75+1)=0.875
t(x|c) = count(x|c) / (count(x|c)+count(y|c))= 0.75 / (0.75+0.25) = 0.75
t(y|c) = count(y|c) / (count(x|c)+count(y|c))= 0.25 / (0.75+0.25) = 0.25

Παράδειγμα (συν)

- Με επανάληψη των βημάτων παίρνω

$t(x|b) = 0.0001$ $t(x|c) = 0.9999$

$t(y|b) = 0.9999$ $t(y|c) = 0.0001$

Η πιθανότητα της διασταυρωμένης ευθυγράμμισης ενισχύθηκε από το δεύτερο ζεύγος προτάσεων (όπου και εκεί ευθυγραμμίζεται το b με το y).

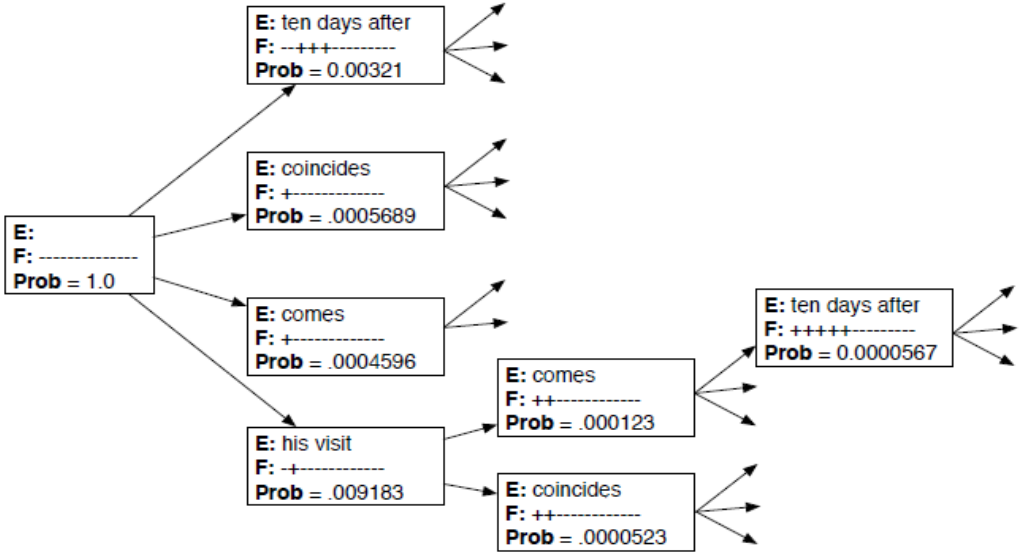
Αυτό ενίσχυσε το $t(y|b)$, και εν συνεχεία και το $t(x|c)$, μια και το x συνδέεται με το c στην ίδια (διασταυρωμένη) ευθυγράμμιση.

Ενίσχυση του $t(x|c)$ σημαίνει υποβάθμιση του $t(y|c)$, μια και αθροίζουν στο 1. Οπότε, παρόλο που τα y και c συνεμφανίζονται, η ανάλυση δείχνει ότι δεν είναι μετάφραση το ένα του άλλου.

Αποκωδικοποίηση - Αναζήτηση

- Βρες τις λέξεις/φράσεις στην γλώσσα-στόχο που μεγιστοποιούν την πιθανότητα του μοντέλου μετάφρασης επί την πιθανότητα του μοντέλου γλώσσας
- Η αναζήτηση (search) πάνω σε όλους τους δυνατούς συνδυασμούς μπορεί να οδηγήσει σε πολύ μεγάλο χώρο αναζήτησης (search space), και πρέπει να βρεθούν τρόποι περιορισμού του
- Ο κόμβος που

καλύπτει όλες τις λέξεις της πρότασης-πηγής και έχει την μεγαλύτερη πιθανότητα κερδίζει.



Στατιστική Μηχανική Μετάφραση: Σύνοψη

- Για να κατασκευάσουμε ένα σύστημα στατιστικής ΜΤ χρειαζόμαστε:
  - Ένα ευθυγραμμισμένο παράλληλο σώμα κειμένων σε δύο γλώσσες (για το μοντέλο μετάφρασης)
  - Ένα μονόγλωσσο σώμα κειμένων για τη κάθε γλώσσα (για το μοντέλο γλώσσας)
  - Προγράμματα εκπαίδευσης (για την εξαγωγή πιθανοτήτων από τα σώματα κειμένων και την εκπαίδευση των μοντέλων)
  - Έναν «αποκωδικοποιητή» (που για δεδομένη είσοδο, θα ψάχνει την έξοδο που αντιστοιχεί στη μεγαλύτερη πιθανότητα)
- Δείτε ένα παράδειγμα με διαθέσιμο software
  - (http://www.clsp.jhu.edu/ws99/projects/mt/)

Μετάφραση Βάσει Παραδειγμάτων

- Χρήση μιας βάσης παραδειγμάτων από προηγούμενες μεταφράσεις για την καθοδήγηση της διαδικασίας μετάφρασης
- Συνώνυμοι όροι:
  - Example-based MT
  - Analogy-based MT
  - Memory-based MT
  - Case-based MT
  - Experience-guided MT

Βασική Ιδέα

- Ταίριαξε κομμάτια του κειμένου με παραδείγματα από τη βάση
- Αναγνώρισε τις μεταφράσεις των κομματιών
- Συνδύασε τις μεταφράσεις των κομματιών για τη δημιουργία του κειμένου μετάφρασης
- Πλεονέκτημα: Η χρήση πραγματικών παραδειγμάτων ανθρώπινων μεταφράσεων μπορεί να οδηγήσει σε πιο ποιοτική μετάφραση
- Μειονέκτημα: Έχει περιορισμένη καλυψιμότητα, ανάλογα με το μέγεθος της βάσης παραδειγμάτων

Παράδειγμα

- Είσοδος:
  - Αγόρασα ένα βιβλίο εξωτερικής πολιτικής
- Παραδείγματα που ταιριάζουν:
  - Αγόρασα ένα καινούργιο αυτοκίνητο

    I bought a new car
  - Διάβασα ένα βιβλίο εξωτερικής πολιτικής

    I read a foreign affairs book
- Αποτέλεσμα:
  - I bought a foreign affairs book

Προβλήματα

- Πόσα παραδείγματα πρέπει να έχουμε;
- Από που θα τα βρούμε;
- Πώς θα τα αποθηκεύσουμε;
- Πώς θα δουλεύει ο αλγόριθμος ταιριάσματος;
  - Εξαρτάται από τον τρόπο αποθήκευσης των παραδειγμάτων
- Πώς επανασυνδέονται τα κομμάτια της μετάφρασης;
- Από τις περισσότερες υποψήφιες μεταφράσεις πώς θα επιλέγεται η καλύτερη;

Τα Παραδείγματα

- Πόσα;
  - Εξαρτάται από το σύστημα
  - Συνήθως μεταξύ 1k και 10k
- Από πού;
  - Εξαγωγή από παράλληλα κείμενα
  - Χειρονακτική κατασκευή
  - Μπορεί να χρειαστεί κάποιο φιλτράρισμα

Επικαλυπτόμενα Παραδείγματα

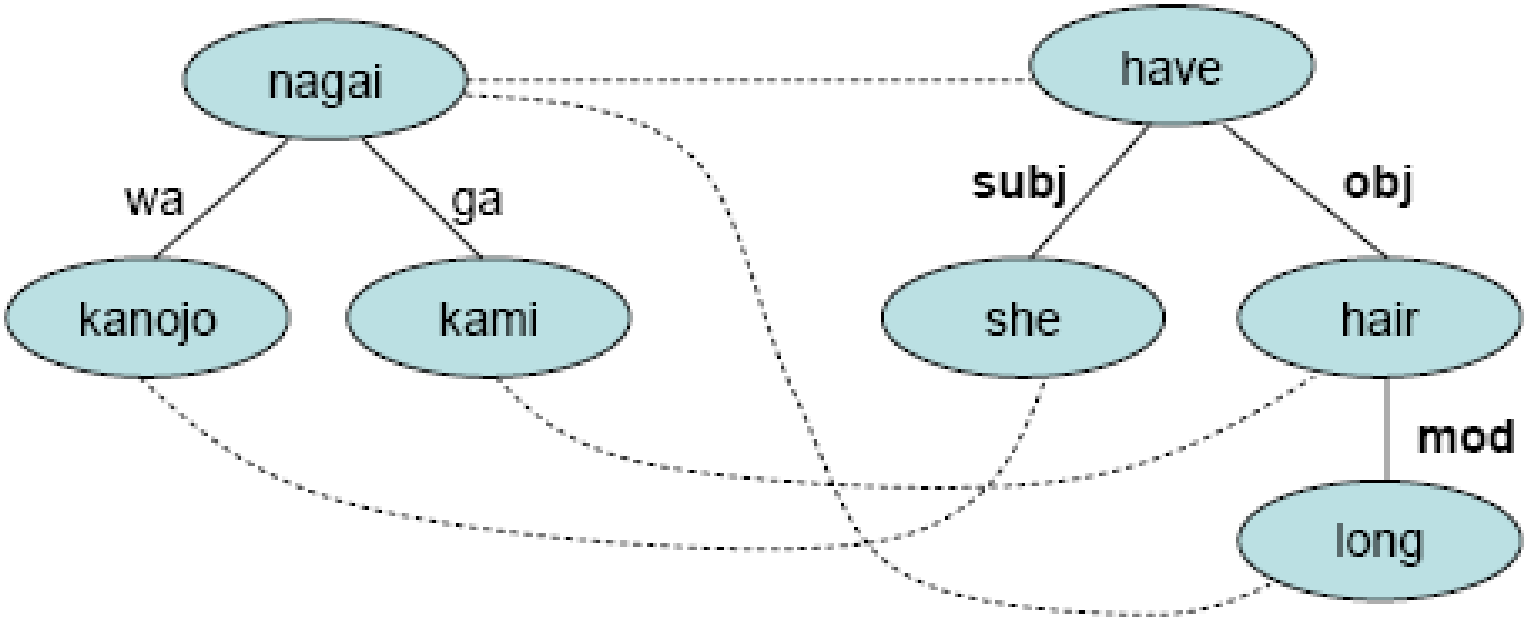
- Όταν δύο παραδείγματα επικαλύπτονται υπάρχουν δύο περιπτώσεις:
  - Είναι ταυτόσημα ή ουσιαστικά ίδια: το ένα ενισχύει το άλλο
  - Είναι αντιφατικά οδηγώντας σε διαφορετικές μεταφράσεις

Αποθήκευση Παραδειγμάτων

- Η απλούστερη περίπτωση είναι η αποθήκευση ζευγών αλφαριθμητικών
- Συχνά τα παραδείγματα είναι σχολιασμένα ως προς το μέρος-του-λόγου ή με κάποια άλλη απλή πληροφορία
- Πιο περίπλοκοι τρόποι:
  - Σχολιασμένες δομές δέντρων
  - Γενικευμένα παραδείγματα

Παράδειγμα: Σχολιασμένες Δομές Δέντρων

- Japanese → English
- J: *Kanojo wa kami ga nagai*
- E: *She has long hair*



Παράδειγμα: Γενικευμένα Παραδείγματα

- Παρόμοια παραδείγματα ομαδοποιούνται:
- John Miller flew to Frankfurt on December 3rd.
- <1stname> <lastname> flew to <city> on <date>.
- <person> flew to <city> on <date>.
- Dr. Howard Johnson flew to Ithaca on 7 April 1997.

Αλγόριθμος Ταιριάσματος

- Βάσει χαρακτήρων (απλή σύγκριση αλφαριθμητικών)
- Βάσει λέξεων (με θησαυρό όρων)
- Βάσει δομής (αν τα παραδείγματα είναι αποθηκευμένα σε δέντρο)
- Μερικό ταίριασμα
  - Εύρεση ενός συνόλου παραδειγμάτων που ταιριάζουν μερικώς (chunks)

Συνδυασμός Κομματιών Μετάφρασης

- Τα μεταφρασμένα παραδείγματα που εξάγονται από τον αλγόριθμο ταιριάσματος πρέπει να συνδεθούν με τρόπο που να είναι νόμιμος στη γλώσσα προορισμού
- Είσοδος:
  - *The tall man entered the room*
- Παραδείγματα που ταιριάζουν:
  - *The tall man ate his breakfast*
  - *Ο ψηλός άντρας έφαγε το πρωινό του*
  - *I saw the tall man*
  - *Είδα τον ψηλό άντρα*

Λύση του προβλήματος

– Στοιχειώδης γραμματική της γλώσσας προορισμού

– Πληροφορία δεξιών – αριστερών συμφραζομένων

Αξιολόγηση της Αυτόματης Μετάφρασης

- Αξιολόγηση συστημάτων ΜΜ
  - Για την ταξινόμηση των συστημάτων
  - Για την αξιολόγηση των αλλαγών που πραγματοποιούνται σε ένα σύστημα με στόχο την βελτίωσή του
- Τρόποι αξιολόγησης
  - Έμμεσα, μέσω άλλων εργασιών
    - Κατανόηση κειμένου
    - Κατασκευή κάποιου πονήματος από κάποιο εγχειρίδιο χρήσης
  - Άμεσα
    - Ευφράδεια (fluency) / επάρκεια (adequacy)
  - Χειρωνακτικά/Αυτόματα

Ευφράδεια (Fluency)

- Κλίμακα 5 σημείων (5 point scale)
  - 5 Άπταιστη γλώσσα
  - 4 Καλή γλώσσα
  - 3 Όχι μητρική γλώσσα
  - 2 Προβληματική γλώσσα
  - 1 Ακατανόητη γλώσσα

Επάρκεια (Adequacy)

- Το μεταφρασμένο κείμενο περιέχει πόση από την πληροφορία που περιέχει το πρότυπο μεταφρασμένο
  - 5 όλη
  - 4 την περισσότερη
  - 3 αρκετή
  - 2 λίγη
  - 1 καθόλου

Χειρωνακτική – Αυτόματη Αξιολόγηση

- Η χειρωνακτική αξιολόγηση
  - Είναι πολύ χρονοβόρα
  - Είναι πολύ ορθή
  - Είναι μη επαναχρησιμοποιήσιμη
- Η αυτόματη αξιολόγηση
  - Είναι φτηνή και επαναχρησιμοποιήσιμη
  - Όχι πάντα αξιόπιστη

Στόχοι αυτόματης αξιολόγησης

- Να δίνει την δυνατότητα κατάταξης των διαφόρων συστημάτων
- Να έχει την δυνατότητα να αναγνωρίζει σε τι είδους προτάσεις δεν τα πάει καλά ένα σύστημα και να κατηγοριοποιεί τα λάθη
- Να παρέχει ένα σκορ που να είναι επεξηγήσιμο
- Να μπορεί να συσχετιστεί με ανθρώπινες κρίσεις

Χρήση «backtranslation»

- Δεν είναι καλή μέθοδος αξιολόγησης

The spirit is willing, but the flesh is weak

English → Russian → English

The vodka is good but the meat is rotten

Μεθοδολογία

- Σύγκριση με πρότυπες μεταφράσεις
- Λογική: όσο πιο κοντά είμαστε σε ανθρώπινες μεταφράσεις, τόσο καλύτερο είναι το σύστημα ΜΜ
- Ποσοστό σφαλμάτων λέξεων (Word Error Rate - WER)
  - Υπολογισμός του ελάχιστου αριθμού εισαγωγών, διαγραφών και αντικαταστάσεων που πρέπει να πραγματοποιηθούν ώστε να μετατραπεί το αυτόματα μεταφρασμένο κείμενο στο κείμενο της πρότυπης μετάφρασης
  - Έχει το πρόβλημα ότι στην ΜΜ υπάρχουν πολλοί πιθανοί και εξίσου δόκιμοι τρόποι μετάφρασης μιας πρότασης
  - Λύση: χρήση πολλών πρότυπων μεταφράσεων

BiLingual Evaluation Understudy (BLEU)

- Χρησιμοποιεί πολλαπλές πρότυπες μεταφράσεις
- Ψάχνει για ν-γραμμα που εμφανίζονται οπουδήποτε μέσα στην πρόταση
- Έχει και “brevity penalty”
- Στόχος: να διακρίνει ποιο σύστημα έχει καλύτερη ποιότητα (σε συσχέτιση με άνθρωπους-κριτές)
- Το BLEU μετράει την επικάλυψη με τις πρότυπες μεταφράσεις

Bleu - Παράδειγμα

R1: It is a guide to action that ensures that the military will forever heed Party commands.
R2: It is the Guiding Principle which guarantees the military forces always being under the command of the Party.
R3: It is the practical guide for the army always to heed the directions of the party.

C1: It is to insure the troops forever hearing the activity guidebook that party direct.
C2: It is a guide to action which ensures that the military always obeys the command of the party.

Ταίριασμα ν-γράμμων με την πρώτη μετάφραση

R1: It is a guide to action that ensures that the military will forever heed Party commands.
R2: It is the Guiding Principle which guarantees the military forces always being under the command of the Party.
R3: It is the practical guide for the army always to heed the directions of the party.

C1: It is to insure the troops forever hearing the activity guidebook that party direct.

Ταίριασμα ν-γράμμων με την δεύτερη μετάφραση

R1: It is a guide to action that ensures that the military will forever heed Party commands.
R2: It is the Guiding Principle which guarantees the military forces always being under the command of the Party.
R3: It is the practical guide for the army always to heed the directions of the party.

C2: It is a guide to action which ensures that the military always obeys the command of the party.

- Επειδή η C2 έχει περισσότερα ν-γραμμα και μεγαλύτερα ν-γραμμα από την C1 παίρνει μεγαλύτερο σκορ
- Έχει βρεθεί ότι το Bleu συσχετίζεται με την ανθρώπινη κρίση ποιότητας μετάφρασης
- Πώς εξηγείται το σκορ;
  - Πόσα λάθη έχουν γίνει;
  - Πόσο καλύτερο είναι ένα σύστημα σε σχέση με ένα άλλο;
  - Πόσο χρήσιμο είναι το σύστημα;
  - Πόσο πρέπει να βελτιωθεί για να γίνει χρήσιμο;
  - Πόσο καλά συσχετίζεται με ανθρώπινες κρίσεις;

Euromatrix

- Πρακτικά του Ευρωκοινοβουλίου

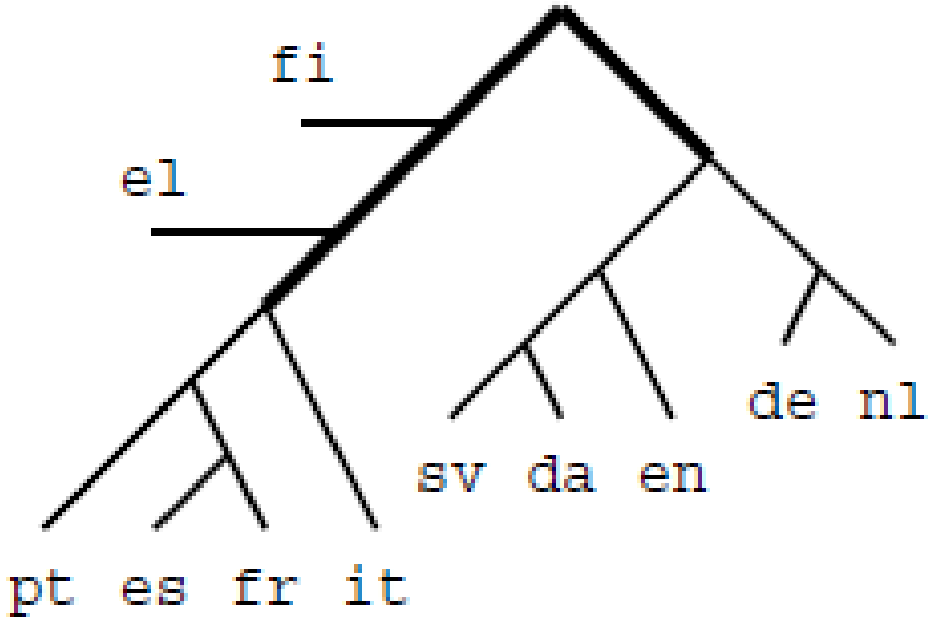
- Μεταφρασμένα σε 11 επίσημες γλώσσες
- Εμπλουτίζεται συνεχώς
- Europarl corpus
  - 20-30 εκατ λέξεις/γλώσσα
  - 110 ζεύγη γλωσσών - 110 συστήματα μετάφρασης

- Scores for all 110 systems

	da	de	el	en	es	fr	fi	it	nl	pt	sv
da	-	18.4	21.1	28.5	26.4	28.7	14.2	22.2	21.4	24.3	28.3
de	22.3	-	20.7	25.3	25.4	27.7	11.8	21.3	23.4	23.2	20.5
el	22.7	17.4	-	27.2	31.2	32.1	11.4	26.8	20.0	27.6	21.2
en	25.2	17.6	23.2	-	30.1	31.1	13.0	25.3	21.0	27.1	24.8
es	24.1	18.2	28.3	30.5	-	40.2	12.5	32.3	21.4	35.9	23.9
fr	23.7	18.5	26.1	30.0	38.4	-	12.6	32.4	21.1	35.3	22.6
fi	20.0	14.5	18.2	21.8	21.1	22.4	-	18.3	17.0	19.1	18.8
it	21.4	16.9	24.8	27.8	34.0	36.0	11.0	-	20.0	31.2	20.2
nl	20.5	18.3	17.4	23.0	22.9	24.6	10.3	20.0	-	20.7	19.0
pt	23.2	18.2	26.4	30.1	37.9	39.0	11.9	32.0	20.2	-	21.9
sv	30.3	18.9	22.8	30.2	28.6	29.7	15.3	23.9	21.9	25.9	-

[from Koehn, 2005: Europarl]

Ομαδοποίηση γλωσσών ανάλογα με το πόσο εύκολα μεταφράζονται η μια στην άλλη – Προσέγγιση των οικογενειών των γλωσσών



[from Koehn, 2005, MT Summit]

Μετάφραση από και προς σε μια γλώσσα

- Κάποιες γλώσσες είναι πιο εύκολο να αποτελούν στόχο, παρά πηγή.
- Οι μορφολογικά πλούσιες γλώσσες (πχ Γερμανικά, Φινλανδικά) είναι πιο δύσκολο να παραχθούν)

Language	From	Into	Diff
da	23.4	23.3	0.0
de	**22.2**	**17.7**	**-4.5**
el	23.8	22.9	-0.9
en	**23.8**	**27.4**	**+3.6**
es	26.7	29.6	+2.9
fr	26.1	31.1	+5.1
fi	19.1	12.4	-6.7
it	24.3	25.4	+1.1
nl	19.7	20.7	+1.1
pt	26.1	27.0	+0.9
sv	24.8	22.1	-2.6

[from Koehn, 2005: Europarl]