Επεξεργασία Φυσικής Γλώσσας & Μηχανική Μάθηση

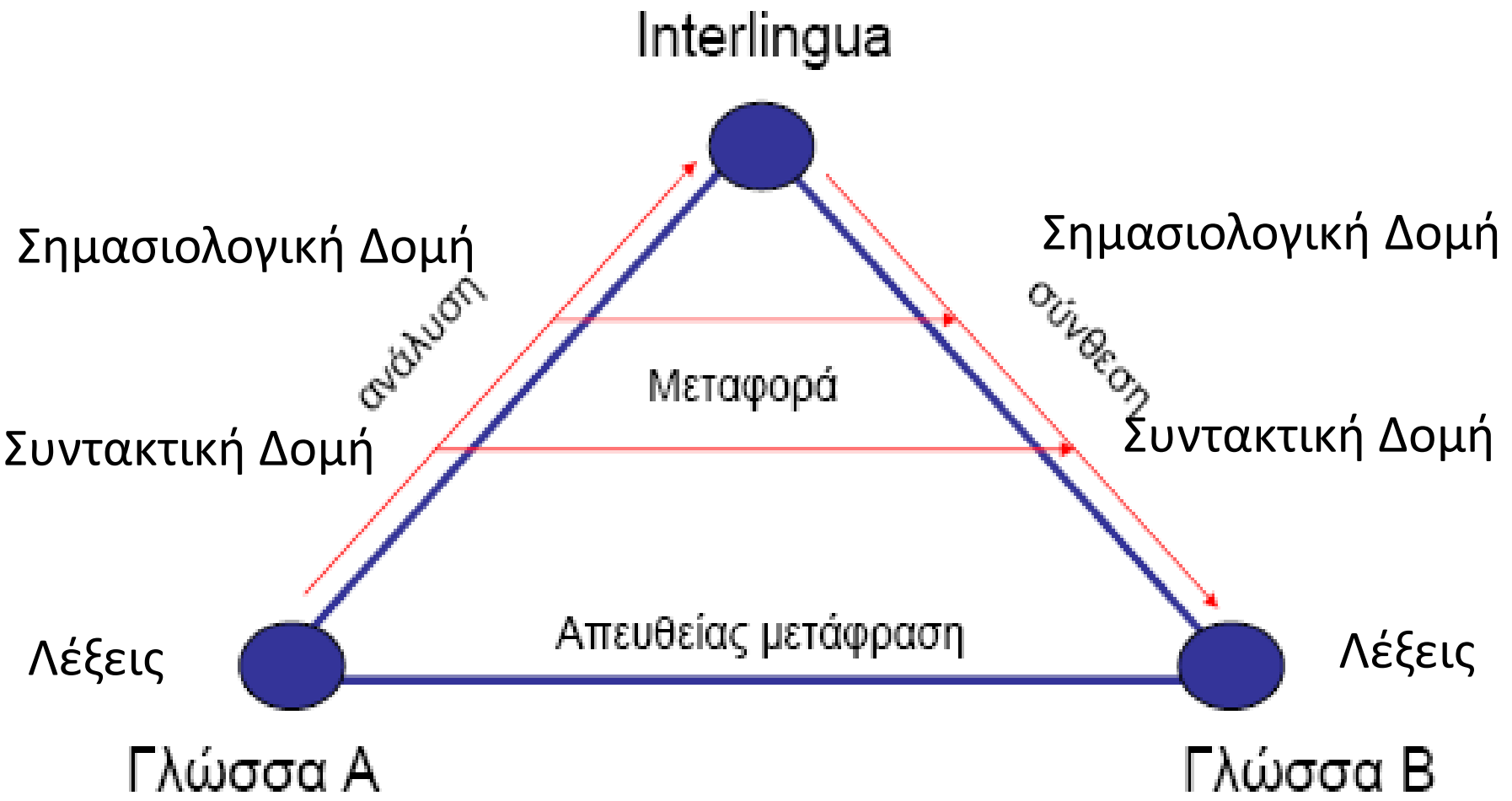
Μηχανική Μετάφραση

Κάτια Κερμανίδου
kerman@ionio.gr

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



Στατιστική ΜΜ

• 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)

$P(B|A) = P(B) * P(A|B) / P(A)$

$B = \text{argmax}_B P(B)*P(A|B)$

P(B): μοντέλο γλώσσας - **το είδαμε στην διάλεξη 2!**

P(A|B): μοντέλο μετάφρασης

Δηλ. την πρόταση B που μεγιστοποιεί την p(B|A)

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

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

Μοντέλο Μετάφρασης (Translation model)

- 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)

Φαινόμενο 1: Μετάφραση

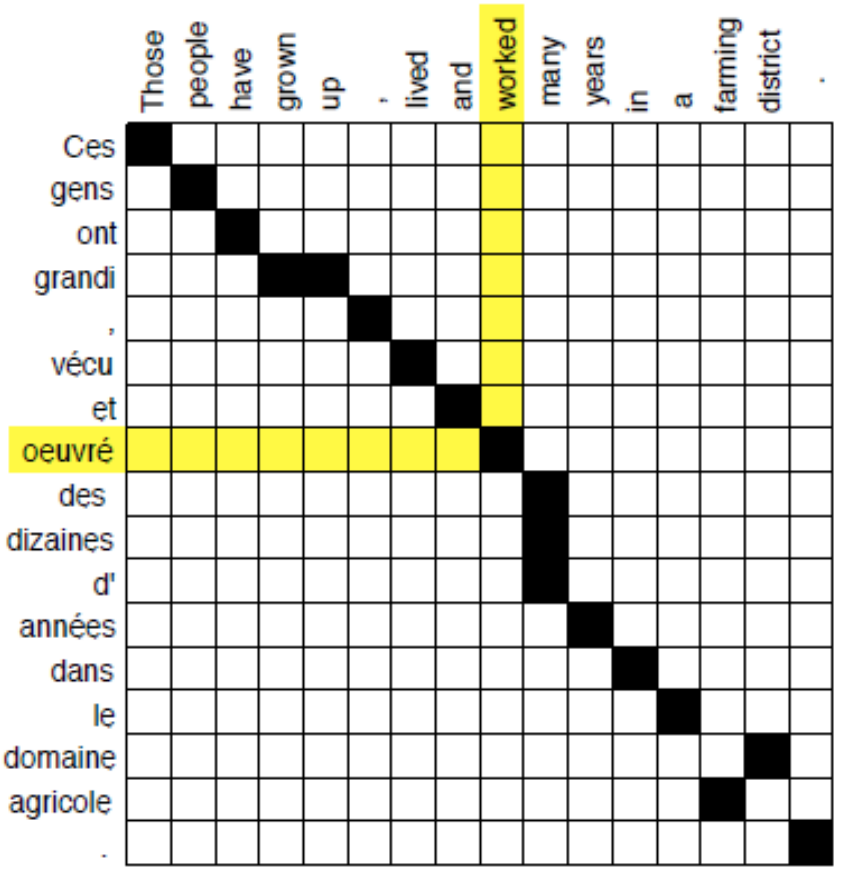
- Πρέπει για το κάθε υπο-συστατικό (λέξη ή φράση) να βρεθεί η μετάφρασή του ανάμεσα στα ζευγάρια προτάσεων του παράλληλου σώματος
- Δεδομένης της λέξης *worked*

ποια η πιθανότητα να έχει

Προκύψει από μετάφραση της

λέξης *travaille? marche?*

oeuvre? fonctionne? …



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

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

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

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

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

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

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.**

Φαινόμενο 2: Γονιμότητα - Fertility

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

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

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

f(Bi): η γονιμότητα της λέξης *Bi*

Φαινόμενο 3: Παραμόρφωση ή Στρέβλωση (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:

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

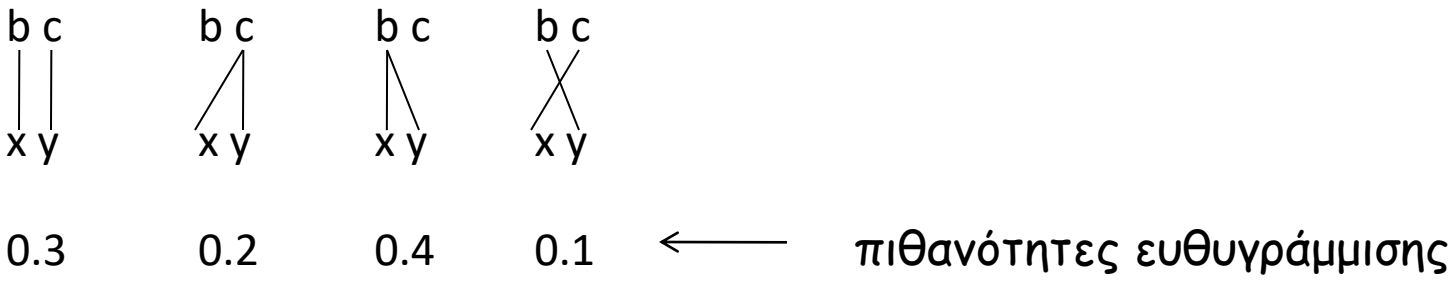
$d(j \mid a_j, l, m)$: η πιθανότητα η λέξη στην θέση j στην πρόταση-πηγή να μεταφραστεί σε μια λέξη που θα βρίσκεται στην θέση a_j στην γλώσσα στόχο.

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 με χρήση των **fractional counts**. Πχ

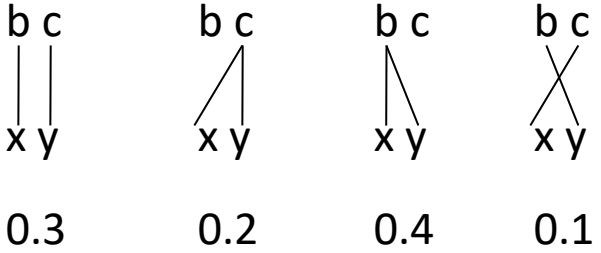


Η πιθανότητα η λέξη b να παράγει ακριβώς μια λέξη ως μετάφρασή της

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 ΕΜ (Expectation Maximization)

- Αλγόριθμος μη επιβλεπόμενης μάθησης
- 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

ΕΜ: Παράδειγμα

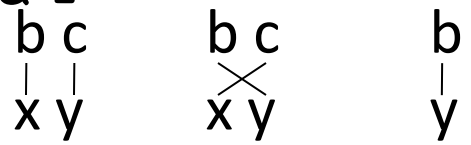
- Έστω το σώμα κειμένων

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. Κανονικοποίηση

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*0.5=0.375

p(α, A|B)=0.75

Παράδειγμα (συν)

Βήμα 6. Κανονικοποίηση

p(a| A,B)=0.125/(0.125+0.375)=0.25

p(a| A,B)=0.375/(0.125+0.375)=0.75

p(a| 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).

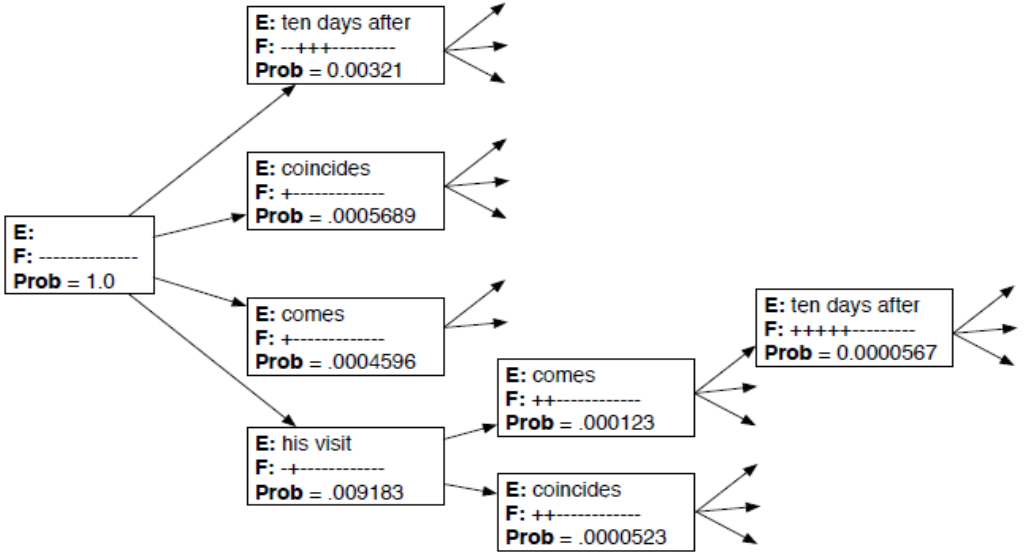
Αυτό ενίσχυσε το $t(y|b)$, και εν συνεχεία και το $t(x|c)$, μια και το x συνδέεται με το c στην ίδια (διασταυρωμένη) ευθυγράμμιση.

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 τ
- Ο κόμβος που

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



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
    - Χειρωνακτικά/Αυτόματα
      - Ευφράδεια (fluency) / επάρκεια (adequacy)

Χειρωνακτική Αξιολόγηση: Ευφράδεια (Fluency)

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

Χειρωνακτική Αξιολόγηση: 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

Αυτόματη αξιολόγηση: Στόχοι

- 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

Αυτόματη αξιολόγηση: Word Error Rate

- 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.

Σύγκριση

- 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

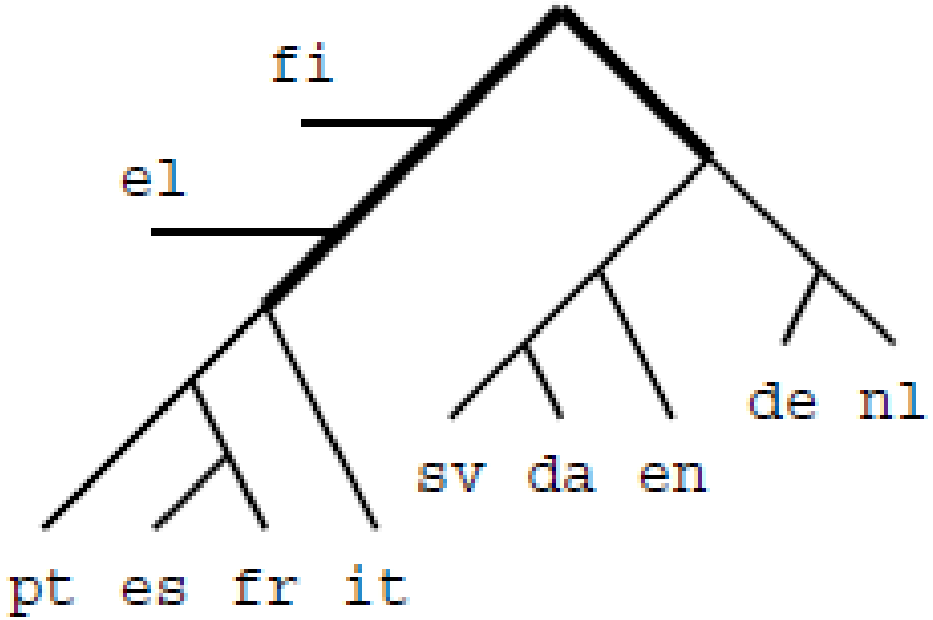
- Πρακτικά του Ευρωκοινοβουλίου
- Μεταφρασμένα σε 11 επίσημες γλώσσες
- Εμπλουτίζεται συνεχώς
- Europarl corpus
  - 20-30 εκατ λέξεις/γλώσσα
  - 110 ζεύγη γλωσσών - 110 συστήματα μετάφρασης

- 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]