ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΩΝ

ΤΜΗΜΑ ΜΗΧ/ΚΩΝ Η/Υ & ΠΛΗΡΟΦΟΡΙΚΗΣ

ΤΕΧΝΗΤΗ ΝΟΗΜΟΣΥΝΗ

Επεξεργασία Φυσικής Γλώσσας

Β. Μεγαλοοικονόμου

(βασισμένο στις διαφάνειες του Γ. Ρεφανίδη, Καθηγητή, Πανεπιστημίου Μακεδονίας και στο βιβλίο «Τεχνητή νοημοσύνη: Μια σύγχρονη προσέγγιση» (4η αμερικανική έκδοση) των Russell & Norvig)

Εισαγωγή

- Η δοκιμασία Turing για τη νοημοσύνη βασίστηκε στη γλώσσα.
- Ένας ομιλητής (ή συντάκτης) έχει τον στόχο να μεταδώσει γνώση, οπότε σχεδιάζει τη γλώσσα που θα αναπαραστήσει αυτήν τη γνώση, και ενεργεί για την επίτευξη του στόχου του. Ο ακροατής (ή αναγνώστης) αντιλαμβάνεται τη γλώσσα, και συμπεραίνει την επιδιωκόμενη σημασία.
- Τρεις βασικοί λόγοι για τους οποίους οι υπολογιστές πρέπει να πραγματοποιούν επεξεργασία φυσικής γλώσσας (natural language processing, NLP):
  - Για να επικοινωνούν (communicate) με τους ανθρώπους.
  - Για να μαθαίνουν (learn).
  - Για να συμβάλλουν στην πρόοδο της επιστημονικής κατανόησης (scientific understanding) των γλωσσών και της χρήσης τους.

Τυπικές έναντι φυσικών γλωσσών

- Οι **τυπικές γλώσσες** (formal languages), όπως η λογική πρώτης τάξης, ορίζονται με ακρίβεια μέσω της **γραμματικής** τους (grammar) που ορίζει τη σύνταξη έγκυρων προτάσεων και των **σημασιολογικών κανόνων** (semantic rules) που ορίζουν τη σημασία τους.
- Οι **φυσικές γλώσσες** δεν μπορούν να οριστούν με τέτοια σαφήνεια.
  - Οι γλωσσικές επιλογές ποικίλλουν από άτομο σε άτομο και από καιρό σε καιρό.
  - Η φυσική γλώσσα είναι **αμφίσημη** και **ασαφής**.
  - Η αντιστοίχιση συμβόλων σε αντικείμενα δεν ορίζεται με τυπικό τρόπο.

Περιπλοκότητες της πραγματικής φυσικής γλώσσας

Πολυπλοκότητα (1/2)

- Έστω η πρόταση «Every agent feels a breeze» (Κάθε πράκτορας αισθάνεται μια αύρα).
- Η πρόταση έχει μία μόνο συντακτική ανάλυση με βάση τη γραμματική της $\mathcal{E}0$, αλλά σημασιολογικά είναι αμφίσημη:
  - –υπάρχει μία αύρα που γίνεται αισθητή από όλους τους πράκτορες, ή μήπως
  - –κάθε πράκτορας αισθάνεται μια ξεχωριστή αύρα που δεν την αισθάνονται οι άλλοι;
- Οι δύο ερμηνείες μπορούν να αναπαρασταθούν ως εξής:

$$\forall a \;\; a \in \Pi\rho\acute{\alpha}\kappa\tau o\rho\varepsilon\varsigma \Rightarrow \exists b \;\; b \in A\acute{\upsilon}\rho\varepsilon\varsigma \wedge A\iota\sigma\theta\acute{\alpha}\nu\varepsilon\tau\alpha\iota(a,b)\,;$$

$$\exists b \;\; b \in A\acute{\upsilon}\rho\varepsilon\varsigma \wedge \forall a \;\; a \in \Pi\rho\acute{\alpha}\kappa\tau o\rho\varepsilon\varsigma \Rightarrow A\iota\sigma\theta\acute{\alpha}\nu\varepsilon\tau\alpha\iota(a,b)\,.$$

Πολυπλοκότητα (2/2)

- Μια καθιερωμένη προσέγγιση στην ποσοτικοποίηση (quantification) είναι η γραμματική να ορίζει όχι μια πραγματική λογική σημασιολογική πρόταση, αλλά μια **ημιλογική μορφή** (quasi-logical form) που στη συνέχεια μετατρέπεται σε λογική πρόταση από αλγορίθμους εκτός της διαδικασίας της συντακτικής ανάλυσης.
- Αυτοί οι αλγόριθμοι μπορεί να έχουν κανόνες προτίμησης για την επιλογή ενός εύρους ποσοδεικτών έναντι άλλων –προτιμήσεις που δεν χρειάζεται να αντανακλώνται άμεσα στη γραμματική.

Πραγματολογία (1/2)

- Η πραγματολογία αφορά το πρόβλημα της συμπλήρωσης της ερμηνείας με την προσθήκη πληροφοριών εξαρτώμενων από τα συμφραζόμενα και σχετικών με την τρέχουσα κατάσταση.
- Η πιο εμφανής ανάγκη για πραγματολογικές πληροφορίες αφορά τον προσδιορισμό της σημασίας των ενδεικτών (indexicals), οι οποίοι είναι φράσεις που αναφέρονται άμεσα στην τρέχουσα κατάσταση.
  - –Για παράδειγμα, στην πρόταση «Εγώ σήμερα είμαι στη Βοστώνη», τόσο το «Εγώ» όσο και το «σήμερα» αποτελούν ενδείκτες.

Πραγματολογία (2/2)

- Ένα άλλο κομμάτι της πραγματολογίας είναι η ερμηνεία της πρόθεσης του ομιλητή.
- Η διατύπωση του ομιλητή θεωρείται γλωσσική πράξη (speech act) και η αποκρυπτογράφηση του τύπου της –ερώτηση, ισχυρισμός, υπόσχεση, προειδοποίηση, εντολή, κ.ο.κ.–επαφίεται στον ακροατή.
- Πρέπει να επεκτείνουμε τη γραμματική ώστε να καλύπτει και τις άλλες μορφές προτάσεων, όπως ερωτήσεις, εντολές, κλπ.
  - –Εντολή ονομάζεται μια ρηματική φράση της οποίας ως υποκείμενο εννοείται ο αποδέκτης (ακροατής) της εντολής:

$$Π(Εντολή(κατηγόρημα(Ακροατής))) \rightarrow PΦ(κατηγόρημα)$$

Εξαρτήσεις μεγάλης απόστασης

- Στην Εικόνα 23.8 είδαμε ότι η φράση «she didn’t hear or even see him», που αναλύθηκε συντακτικά με δύο κενά λόγω της έλλειψης μιας ΟΦ, αναφέρεται ωστόσο στην ΟΦ«him».
  - Γενικά, η απόσταση ανάμεσα στο κενό και την ΟΦ στην οποία αναφέρεται μπορεί να είναι αυθαίρετα μεγάλη.
- Χρησιμοποιείται ένα πολύπλοκο σύστημα επαυξημένων κανόνων για να εξασφαλιστεί ότι οι ΟΦ που λείπουν θα ταιριάζουν στην πρόταση.
  - Π.χ., «What did you [ΡΦ [ΡΦ smell _] και [ΡΦ shoot an arrow at _]]?»

Χρόνος και ρηματικοί χρόνοι

- Θέλουμε να μπορούμε να αναπαραστήσουμε τη διαφορά μεταξύ του «Ali loves Bo» και του «Ali loved Bo».
- Μια καλή επιλογή για την αναπαράσταση του χρόνου των συμβάντων είναι η σημειογραφία του λογισμού συμβάντων:

$$\text{Ali loves Bo: } E_1 \in Αγαπά(Ali, Bo) \wedge ΣτηΔιάρκεια(Τώρα, Μέχρι(E_1))$$
$$\text{Ali loved Bo: } E_2 \in Αγαπά(Ali, Bo) \wedge Μετά(Τώρα, Μέχρι(E_2))$$

- Αυτό υποδηλώνει ότι οι δύο λεξικολογικοί κανόνες για τις λέξεις «loves» και «loved» θα πρέπει να είναι οι εξής:

$$Ρήμα(\lambda y\ \lambda x\ e \in Αγαπά(x, y) \wedge ΣτηΔιάρκεια(Τώρα, e)) \rightarrow \textbf{loves}$$
$$Ρήμα(\lambda y\ \lambda x\ e \in Αγαπά(x, y) \wedge Μετά(Τώρα, e)) \rightarrow \textbf{loved}$$

Αμφισημία (1/2)

- ■Όταν οι ερευνητές ξεκίνησαν να χρησιμοποιούν υπολογιστές για την ανάλυση της γλώσσας στη δεκαετία του 1960, έμειναν έκπληκτοι από το γεγονός ότι σχεδόν όλες οι προτάσεις είναι αμφίσημες, με πολλές δυνατές συντακτικές αναλύσεις (μερικές φορές εκατοντάδες), ακόμα και όταν οι φυσικοί ομιλητές παρατηρούν μόνο μία προτιμώμενη ανάλυση.
  - –Κατανοούμε τη φράση «καστανή ζάχαρη και κόκκινη μπύρα» ως «[καστανή ζάχαρη] και [κόκκινη μπύρα]», και δεν εξετάζουμε ποτέ την ελάχιστα πιθανή ερμηνεία «καστανή [ζάχαρη και κόκκινη μπύρα]», όπου το επίθετο «καστανή» τροποποιεί ολόκληρη τη φράση και όχι μόνο τη «ζάχαρη».
  - –Ομοίως, όταν ακούμε τη φράση «Outside of a dog, a book is a person’s bestfriend», ερμηνεύουμε το «outside of» ως «except for», ενώ βρίσκουμε αστείο το λογοπαίγνιο του Groucho Marx «Inside of a dog it’s too dark to read».

Αμφισημία (2/2)

- Η **λεξικολογική αμφισημία** (lexical ambiguity) υπάρχει όταν μια λέξη έχει περισσότερες από μία σημασίες.
  - –Το «back» μπορεί να είναι επίρρημα (go back), επίθετο (back door), ουσιαστικό (the back of the room), ρήμα (back a candidate), ή κύριο όνομα (ποτάμι στην περιοχή Νουναβούτ του Καναδά).
- Η **συντακτική αμφισημία** (syntactic ambiguity) αναφέρεται σε μια φράση που έχει πολλές συντακτικές αναλύσεις.
- Η **συντακτική αμφισημία** οδηγεί σε σημασιολογική αμφισημία (semantic ambiguity).

Μετωνυμία (1/2)

- Η μετωνυμία (metonymy) είναι ένα σχήμα λόγου στο οποίο ένα αντικείμενο χρησιμοποιείται στη θέση κάποιου άλλου.
  - –Όταν ακούμε την πρόταση «Chrysler announced a new model», δεν δίνουμε την ερμηνεία ότι οι εταιρείες μπορούν να μιλούν.
- Για να χειριστούμε σωστά τη σημασιολογία της μετωνυμίας, πρέπει να εισαγάγουμε ένα εντελώς νέο επίπεδο αμφισημίας.
  - –Αυτό μπορούμε να το κάνουμε παρέχοντας δύο αντικείμενα για τη σημασιολογική ερμηνεία κάθε φράσης της πρότασης: το αντικείμενο που εννοείται κυριολεκτικά στη φράση (Chrysler) και αυτό της μετωνυμικής αναφοράς (ο εκπρόσωπος της εταιρείας).

Μετωνυμία (2/2)

- Στην τρέχουσα γραμματική μας, η φράση «Chrysler announced» ερμηνεύεται ως:

$$x = Chrysler \wedge e \in Ανακοινώνει(x) \wedge Μετά(Τώρα, Μέχρι(e))$$

- Πρέπει να αλλάξουμε το παραπάνω σε:

$$x = Chrysler \wedge e \in Ανακοινώνει(m) \wedge Μετά(Τώρα, Μέχρι(e)) \wedge Μετωνυμία(m,x)$$

- Το επόμενο βή ν να προκύψουν. Η απλούστερη περίπτωση είναι να μην υπάρχει καθόλου μετωνυμία:

$$\forall m,x \ \ (m = x) \Rightarrow Μετωνυμία(m,x)$$

- ή να υπάρχει:

$$\forall m,x \ \ x \in Οργανισμοί \wedge Εκπρόσωπος(m,x) \Rightarrow Μετωνυμία(m,x)$$

Άλλες μετωνυμίες περιλαμβάνουν τον συγγραφέα αντί των έργων (I readShakespeare), τον παραγωγό αντί του προϊόντος (I drivea Honda), και το μέρος αντί του συνόλου (The Red Soxneeda strongarm).

Μεταφορά

- Η μεταφορά (metaphor) είναι ένα σχήμα λόγου στο οποίο μια φράση με κυριολεκτικό νόημα χρησιμοποιείται για να υποδηλώσει ένα διαφορετικό νόημα σύμφωνα με κάποια αναλογία.
- Συνεπώς, η μεταφορά μπορεί να θεωρηθεί ως ένα είδος μετωνυμίας όπου η σχέση μεταξύ κυριολεκτικής και μεταφορικής φράσης διέπεται από ομοιότητα.

Αποσαφήνιση

- **Αποσαφήνιση** (disambiguation) αποκαλείται η διαδικασία ανάκτησης της πιο πιθανής επιδιωκόμενης σημασίας μιας διατύπωσης.
- Για να αποσαφηνίζουμε σωστά τον λόγο, θα πρέπει να συνδυάζουμε τέσσερα μοντέλα:
  - –Το **μοντέλο του κόσμου** (world model): Η πιθανότητα να προκύψει μια πρόταση στον κόσμο.
  - –Το **νοητικό μοντέλο** (mental model): Η πιθανότητα ο ομιλητής να σχηματίσει την πρόθεση να γνωστοποιήσει στον ακροατή ένα συγκεκριμένο γεγονός.
  - –Το **γλωσσικό μοντέλο** (language model): Η πιθανότητα να επιλεγεί μια συγκεκριμένη συμβολοσειρά λέξεων, με δεδομένο το νοητικό μοντέλο.
  - –Το **ακουστικό μοντέλο** (acoustic model): Η πιθανότητα να παραχθεί μια συγκεκριμένη ακολουθία ήχων κατά την προφορική επικοινωνία, με δεδομένα τα προηγούμενα μοντέλα.

Εργασίες φυσικής γλώσσας

Αναγνώριση ομιλίας

- Η αναγνώριση ομιλίας (speech recognition) αφορά τον μετασχηματισμό του εκφωνούμενου λόγου σε κείμενο.
- Τα σύγχρονα συστήματα έχουν ποσοστό λανθασμένων λέξεων περίπου 3% έως 5% (ανάλογα με τις λεπτομέρειες του συνόλου ελέγχου), παρόμοιο με αυτό των επαγγελματιών απομαγνητοφωνητών.
- Τα κορυφαία σύγχρονα συστήματα χρησιμοποιούν έναν συνδυασμό αναδρομικών νευρωνικών δικτύων και κρυφών μοντέλων Markov.
- Τα βαθιά νευρωνικά δίκτυα λειτουργούν καλά στο πρόβλημα της αναγνώρισης ομιλίας λόγω της φυσικής διάκρισής της σε επιμέρους συνθετικά: τον μετασχηματισμό κυματομορφών σε φθόγγους, φθόγγων σε λέξεις, και λέξεων σε προτάσεις.

Σύνθεση κειμένου σε ομιλία

- Η σύνθεση κειμένου σε ομιλία (text-to-speech) είναι η αντίστροφη διαδικασία –ο μετασχηματισμός κειμένου σε ήχο.
- Η πρόκληση είναι να προφέρουμε σωστά κάθε λέξη, και να κάνουμε τη ροή κάθε πρότασης να φαίνεται φυσική, με τον σωστό επιτονισμό και τις ανάλογες παύσεις.
- Ο τομέας ασχολείται επίσης με τη σύνθεση διαφορετικών φωνών, τοπικές διαλέκτους, καθώς επίσης και φωνές συγκεκριμένων προσώπων.
- Η υιοθέτηση βαθιών αναδρομικών νευρωνικών δικτύων οδήγησε σε μεγάλη βελτίωση.

Μηχανική μετάφραση

- Η **μηχανική μετάφραση** (machine translation) μετασχηματίζει κείμενο από μια γλώσσα σε μια άλλη.
- Τα συστήματα αυτά εκπαιδεύονται συνήθως με τη χρήση ενός δίγλωσσου σώματος κειμένων: ένα σύνολο ζευγών εγγράφων, εκ των οποίων το ένα είναι, για παράδειγμα, στα Αγγλικά και το άλλο στα Γαλλικά.
- Τα έγγραφα δεν χρειάζεται να επισημειώνονται με κάποιον τρόπο· το σύστημα μηχανικής μετάφρασης μαθαίνει να ευθυγραμμίζει προτάσεις και φράσεις, και όταν παρουσιάζεται μια νέα πρόταση στη μία γλώσσα, τότε μπορεί να παράγει μια μετάφραση στην άλλη.
- Τα τελευταία χρόνια χρησιμοποιήθηκαν κυρίως αναδρομικά νευρωνικά μοντέλα ακολουθίας σε ακολουθία (recurrent neural sequence-to-sequence models), καθώς και νευρωνικά μοντέλα μετασχηματιστή (transformer model)

Εξαγωγή πληροφοριών

- Η **εξαγωγή πληροφοριών** (information extraction) είναι η διαδικασία απόκτησης γνώσης μέσω της σάρωσης ενός κειμένου και της αναζήτησης εμφανίσεων συγκεκριμένων κατηγοριών αντικειμένων, καθώς και των μεταξύ τους σχέσεων.
- Αν το κείμενο προέλευσης είναι καλά δομημένο, τότε οι πληροφορίες μπορούν να εξάγονται με απλές τεχνικές, όπως **κανονικές παραστάσεις** (regular expressions)
- Για κείμενα ελεύθερης μορφής, οι τεχνικές περιλαμβάνουν κρυφά μοντέλα Markovκαι συστήματα μάθησης βασισμένα σε κανόνες.
- Τα πιο πρόσφατα συστήματα, εκμεταλλευόμενα την ευελιξία που παρέχουν οι ενσωματώσεις λέξεων, χρησιμοποιούν αναδρομικά νευρωνικά δίκτυα.

Ανάκτηση πληροφοριών

- **Ανάκτηση πληροφοριών** (information retrieval, IR) είναι η εργασία της εύρεσης εγγράφων που σχετίζονται με ένα δεδομένο ερώτημα και είναι σημαντικά για αυτό.–Διαδικτυακές μηχανές αναζήτησης
- Η **απάντηση ερωτήσεων** (question answering) είναι μια διαφορετική εργασία, στην οποία το ερώτημα αποτελεί πραγματικά ερώτηση, όπως «Ποιος ίδρυσε την Αμερικανική Ακτοφυλακή;», και η απάντηση δεν είναι μια ταξινομημένη λίστα εγγράφων αλλά μια πραγματική απάντηση: «Ο Alexander Hamilton».

Γλωσσικά μοντέλα

- Ορίζουμε ένα **γλωσσικό μοντέλο** (language model) ως μια κατανομή πιθανοτήτων που περιγράφει την πιθανοφάνεια μιας συμβολοσειράς.
    - Τα γλωσσικά μοντέλα αποτελούν τον πυρήνα ενός μεγάλου εύρους εργασιών φυσικής γλώσσας.

Μοντέλο σάκου λέξεων (1/2)

- Θα ασχοληθούμε με την ταξινόμηση κειμένου σε κατηγορίες, π.χ. οικονομία ή καιρός.
- Με δεδομένη μια πρόταση που αποτελείται από τις λέξεις w1, w2, ..., wN (εν συντομία w1:N,

  ο απλοϊκός τύπος Bayes ορίζει ότι:

$$\mathbf{P}(Κατηγορία \mid w_{1:N}) = \alpha\, \mathbf{P}(Κατηγορία) \prod_j \mathbf{P}(w_j \mid Κατηγορία)$$

- Η εφαρμογή ενός απλοϊκού μοντέλου Bayes σε ακολουθίες λέξεων ονομάζεται **μοντέλο σάκου λέξεων** (bag-of-words model).
  - Το μοντέλο είναι σαφώς εσφαλμένο
- Μπορούμε να μάθουμε τις εκ των προτέρων πιθανότητες που απαιτούνται για αυτό το μοντέλο μέσω επιβλεπόμενης εκπαίδευσης σε ένα σώμα κειμένων (corpus), όπου κάθε τμήμα του επισημαίνεται με μια κατηγορία.

Μοντέλο σάκου λέξεων(2/2)

- Μερικές φορές, μια διαφορετική προσέγγιση μηχανικής μάθησης, όπως η λογιστική παλινδρόμηση, τα νευρωνικά δίκτυα ή οι μηχανές διανυσμάτων υποστήριξης (SVM), μπορεί να λειτουργεί ακόμα καλύτερα από τα απλοϊκά δίκτυα Bayes.
  - Τα χαρακτηριστικά του μοντέλου μηχανικής μάθησης είναι οι λέξεις στο λεξιλόγιο και οι τιμές είναι το πλήθος των εμφανίσεων κάθε λέξης στο κείμενο.
- Μερικά μοντέλα μηχανικής μάθησης λειτουργούν καλύτερα όταν κάνουμε **επιλογή χαρακτηριστικών** (feature selection), περιορίζοντας τα χαρακτηριστικά σε ένα υπο-σύνολο των λέξεων.
  - Δεν είναι απλό να αποφασίσουμε τι θεωρείται λέξη (word).
- Η διαδικασία της διαίρεσης ενός κειμένου σε μια ακολουθία λέξεων ονομάζεται **διαχωρισμός σε σύμβολα** (tokenization).

n-γραμμικά μοντέλα λέξεων (1/2)

- Στο αντίποδα του μοντέλου σάκου λέξεων βρίσκεται ένα μοντέλο, όπου η πιθανότητα κάθε λέξης εξαρτάται από όλες τις προηγούμενές της σε μια πρόταση:

$$P(w_{1:N}) = \prod_{j=1}^{N} P(w_j \mid w_{1:j-1})$$

- Μια μέση λύση είναι ένα μοντέλο **αλυσίδας Markov** (Markov chain) που εξετάζει μόνο την εξάρτηση *n* γειτονικών λέξεων. Είναι γνωστό ως *n*-γραμμικό μοντέλο (*n*-gram model – από τη λέξη *γράμμα*): μια ακολουθία γραπτών συμβόλων μήκους *n* ονομάζεται *n*-γραμμική, και ειδικά **«μονογραμμική»,«διγραμμική»** και **«τριγραμμική»** όταν n=1, 2 και 3, αντίστοιχα.

n-γραμμικά μοντέλα λέξεων(2/2)

- Σε ένα *n*-γραμμικό μοντέλο, η πιθανότητα κάθε λέξης εξαρτάται μόνο από τις $n-1$ προηγούμενες:

$$P(w_j \mid w_{1:j-1}) = P(w_j \mid w_{j-n+1:j-1})$$

$$P(w_{1:N}) = \prod_{j=1}^{N} P(w_j \mid w_{j-n+1:j-1})$$

- Τα *n*-γραμμικά μοντέλα λειτουργούν καλά για την ταξινόμηση των ενοτήτων μιας εφημερίδας, καθώς και για άλλες εργασίες ταξινόμησης όπως η **ανίχνευση ανεπιθύμητης αλληλογραφίας** (spam detection), η **ανάλυση συναισθήματος** (sentiment analysis) και η **αναγνώριση συγγραφέα**

Άλλα *n*-γραμμικά μοντέλα

- Μια εναλλακτική των *n*-γραμμικών μοντέλων λέξεων είναι τα **μοντέλα σε επίπεδο χαρακτήρων** (character-level models), όπου η πιθανότητα κάθε χαρακτήρα καθορίζεται από τους προηγούμενους $n-1$ χαρακτήρες.
  - Τα μοντέλα σε επίπεδο χαρακτήρων είναι κατάλληλα για την εργασία της **αναγνώρισης γλώσσας** (language identification).
- Μια άλλη δυνατότητα αποτελεί το μοντέλο **skip-gram** (γραμμικό μοντέλο παράλειψης λέξεων): μετράμε λέξεις που βρίσκονται κοντά η μία στην άλλη, αλλά παραλείπουμε μία ή περισσότερεςλέξεις που παρεμβάλλονται μεταξύ τους.
  - Για παράδειγμα, με δεδομένο το γαλλικό κείμενο «je ne comprends pas», οι διγραμμικές ακολουθίες 1 παράλειψης είναι οι «je comprends» και «ne pas».

Εξομάλυνση *n*-γραμμικών μοντέλων (1/3)

- Τα *n*-γραμμικά χαμηλής συχνότητας έχουν μικρό αριθμό εμφανίσεων που υπόκεινται σε τυχαίο θόρυβο, και άρα υψηλή διακύμανση.
  - Τα μοντέλα μας αποδίδουν καλύτερα αν εξομαλύνουμε αυτήν τη διακύμανση.
- Επιπλέον, υπάρχει πάντα το ενδεχόμενο να μας ζητηθεί να εκτιμήσουμε ένα κείμενο που περιέχει μια άγνωστη λέξη, δηλαδή μια λέξη **εκτός λεξιλογίου** (out-of-vocabulary).
  - Ένας τρόπος μοντελοποίησης άγνωστων (unknown) λέξεων είναι η τροποποίηση του σώματος εκπαίδευσης με την αντικατάσταση των σπάνια εμφανιζόμενων λέξεων με ένα ειδικό σύμβολο, το οποίο παραδοσιακά είναι το σύμβολο <UNK>.
  - Μερικές φορές χρησιμοποιούνται διαφορετικά σύμβολα άγνωστων λέξεων για διαφορετικούς τύπους.

Εξομάλυνση *n*-γραμμικών μοντέλων(2/3)

- Ακόμα και μετά τη διαχείριση των άγνωστων λέξεων, θα έχουμε το πρόβλημα των άγνωστων *n*-γραμμικών ακολουθιών.
- Η **εξομάλυνση Laplace** (Laplace smoothing), ή αλλιώς εξομάλυνση συν ένα (add-one smoothing), προσθέτει +1 στη μέτρηση κάθε *n*-γράμματος και στη συνέχεια διαιρεί όλες τις μετρήσεις διά το άθροισμά τους.
- Μια άλλη επιλογή είναι το **μοντέλο υπαναχώρησης** (back-off model), στο οποίο ξεκινάμε υπολογίζοντας τις μετρήσεις των *n*-γραμμικών ακολουθιών, αλλά για οποιαδήποτε συγκεκριμένη ακολουθία που δίνει χαμηλή (ή μηδενική) μέτρηση, υπαναχωρούμε στις ($n-1$)-γραμμικές ακολουθίες.

Εξομάλυνση *n*-γραμμικών μοντέλων(3/3)

- Η εξομάλυνση **γραμμικής παρεμβολής** (linear interpolation smoothing) είναι ένα μοντέλο υπαναχώρησης που συνδυάζει τριγραμμικά, διγραμμικάκαι μονογραμμικά μοντέλα με γραμμική παρεμβολή. Ορίζουμε την εκτίμηση πιθανότητας ως:

$$\hat{P}(c_i|c_{i-2:i-1}) = \lambda_3 P(c_i|c_{i-2:i-1}) + \lambda_2 P(c_i|c_{i-1}) + \lambda_1 P(c_i)$$

όπου $\lambda 1 + \lambda 2 + \lambda 3 = 1$.

- Οι τιμές των παραμέτρων λ μπορούν να είναι σταθερές ή να εκπαιδευτούν με έναν αλγόριθμο μεγιστοποίησης–αναμονής.
- Είναι επίσης δυνατό οι τιμές των λ να εξαρτώνται από τις μετρήσεις:
  - αν έχουμε μια υψηλή μέτρηση τριγραμμικών μοντέλων, τότε αποδίδουμε σε αυτά σχετικά μεγαλύτερο βάρος
  - αν η μέτρησή τους είναι χαμηλή, αποδίδουμε μεγαλύτερο βάρος στα διγραμμικά και μονογραμμικά μοντέλα.

Αναπαραστάσεις λέξεων

- Τα *n*-γραμμικά μοντέλα είναι ατομικά (atomic): κάθε λέξη διακρίνεται από όλες τις άλλες, αποτελώντας ένα άτομο χωρίς εσωτερική δομή.
  - Ουσιαστικά τα *n*-γραμμικά μοντέλα δεν αξιοποιούν τους κανόνες γραμματικής και συντακτικού.
- Ένας τύπος δομημένου μοντέλου λέξεων είναι το **λεξικό** (dictionary), το οποίο συνήθως κατασκευάζεται με μη αυτόματο τρόπο.
- Για παράδειγμα, το WordNet είναι ένα λεξικό ανοιχτού πηγαίου κώδικα, επιμελημένο από ανθρώπους, σε μορφή αναγνώσιμη από υπολογιστή, που έχει αποδειχθεί χρήσιμο σε πολλές εφαρμογές φυσικής γλώσσας.
- Παράδειγμα καταχώρησης:

```
"kitten" <noun.animal> ("young domestic cat") IS A: young_mammal

"kitten" <verb.body> ("give birth to kittens")
  EXAMPLE: "our cat kittened again this year"
```

Επισημείωση με μέρη του λόγου (1/2)

- Ένας βασικός τρόπος κατηγοριοποίησης των λέξεων είναι με βάση το **μέρος του λόγου** (part of speech, POS) που αποτελούν.
  - Ονομάζεται επίσης και **λεξική κατηγορία** (lexical category) ή **ετικέτα** (tag) και μπορεί να είναι ουσιαστικό, ρήμα, επίθετο, κλπ.
- Τα μέρη του λόγου επιτρέπουν στα γλωσσικά μοντέλα να αποτυπώνουν γενικεύσεις, όπως ότι «στα Ελληνικά, τα επίθετα συνήθως προηγούνται των ουσιαστικών».
- Η εικόνα στην επόμενη σελίδα παρουσιάζει τις 45 ετικέτες που χρησιμοποιούνται στο Penn Treebank, ένα σώμα κειμένων με περισσότερες από τρία εκατομμύρια λέξεις, επισημειωμένες με το κατάλληλο μέρος του λόγου.
- Ακολουθεί ένα απόσπασμα από το Penn Treebank:

From	the	start	,	it	took	a	person	with	great	qualities	to	succeed
IN	DT	NN	,	PRP	VBD	DT	NN	IN	JJ	NNS	TO	VB

Ετικέτα	Λέξη	Περιγραφή	Ετικέτα	Λέξη	Περιγραφή
CC	*and*	Coordinating conjunction	PRP$	*your*	Possessive pronoun
CD	*three*	Cardinal number	RB	*quickly*	Adverb
DT	*the*	Determiner	RBR	*quicker*	Adverb, comparative
EX	*there*	Existential there	RBS	*quickest*	Adverb, superlative
FW	*per se*	Foreign word	RP	*off*	Particle
IN	*of*	Preposition	SYM	+	Symbol
JJ	*purple*	Adjective	TO	*to*	to
JJR	*better*	Adjective, comparative	UH	*eureka*	Interjection
JJS	*best*	Adjective, superlative	VB	*talk*	Verb, base form
LS	*1*	List item marker	VBD	*talked*	Verb, past tense
MD	*should*	Modal	VBG	*talking*	Verb, gerund
NN	*kitten*	Noun, singular or mass	VBN	*talked*	Verb, past participle
NNS	*kittens*	Noun, plural	VBP	*talk*	Verb, non-3rd-sing
NNP	*Ali*	Proper noun, singular	VBZ	*talks*	Verb, 3rd-sing
NNPS	*Fords*	Proper noun, plural	WDT	*which*	Wh-determiner
PDT	*all*	Predeterminer	WP	*who*	Wh-pronoun
POS	*'s*	Possessive ending	WP$	*whose*	Possessive wh-pronoun
PRP	*you*	Personal pronoun	WRB	*where*	Wh-adverb
$	$	Dollar sign	#	#	Pound sign
“	‘	Left quote	”	’	Right quote
(	[	Left parenthesis	)	]	Right parenthesis
,	,	Comma	.	!	Sentence end
:	;	Mid-sentence punctuation			

Επισημείωση με μέρη του λόγου (2/2)

- Η εργασία της αντιστοίχισης ενός μέρους του λόγου σε κάθε λέξη μιας πρότασης ονομάζεται **επισημείωση με μέρη του λόγου** (part-of-speech tagging).
- Μια συνηθισμένη κατηγορία μοντέλων για την επισημείωση κειμένου με μέρη του λόγου είναι τα **κρυφά μοντέλα Markov** (hidden Markov models, **HMM**).
  - Ένα κρυφό μοντέλο Markov λαμβάνει μια χρονική ακολουθία παρατηρήσεων μαρτυρίας και προβλέπει τις πιο πιθανές κρυφές καταστάσεις που θα μπορούσαν να έχουν δημιουργήσει αυτήν την ακολουθία.
- Τα μοντέλα HMM είναι παραγωγικά (παραγωγή προτάσεων), ωστόσο αποτελούν ένα χρήσιμο μοντέλο αν εφαρμόσουμε τον **αλγόριθμο Viterbi** για να βρούμε την πιο πιθανή ακολουθία κρυφών καταστάσεων (ετικέτες).
  - Συνήθως επιτυγχάνουν πολύ μεγάλη ακρίβεια, γύρω στο 97%.

Κρυφά μοντέλα Markov

- Για να δημιουργήσουμε ένα μοντέλο HMM για επισημείωση POS, χρειαζόμαστε το μοντέλο μετάβασης, που δίνει την πιθανότητα ενός μέρους του λόγου μετά από κάποιο άλλο, P(Ct | Ct-1), και το μοντέλο αισθητήρα, P(Wt | Ct).
  - Για παράδειγμα, στην αγγλική γλώσσα ισχύει P(Ct = VB | Ct-1 = MD) ≅ 0,8 και P(Wt = would | Ct = MD) ≅0,1
- Μια αδυναμία των μοντέλων HMM είναι ότι όλα όσα γνωρίζουμε για τη γλώσσα θα πρέπει να εκφράζονται μέσω των μοντέλων μετάβασης και αισθητήρα.
  - Δεν υπάρχει εύκολος τρόπος για έναν προγραμματιστή συστημάτων να πει, για παράδειγμα, ότι οποιαδήποτε λέξη τελειώνει σε «ικός» είναι πιθανότατα επίθετο ή ότι στη φράση «βαρομετρικό χαμηλό» το βαρομετρικό είναι ουσιαστικό, και όχι επίθετο.

Λογιστική παλινδρόμηση (1/4)

- Η επισημείωση POS μπορεί να γίνει με **λογιστική παλινδρόμηση** (logistic regression)
  - Χρειαζόμαστε ένα διάνυσμα τιμών χαρακτηριστικών **x** και ένα προ-εκπαιδευμένο διάνυσμα βαρών **w**.
  - Χρησιμοποιούμε το εσωτερικό γινόμενο, $\mathbf{w} \cdot \mathbf{x}$ και μετασχηματίζουμε αυτό το άθροισμα σε έναν αριθμό μεταξύ 0 και 1 που μπορεί να ερμηνευτεί ως η πιθανότητα να είναι η είσοδος ένα θετικό παράδειγμα μιας κατηγορίας
- Τα βάρη στο μοντέλο λογιστικής παλινδρόμησης αντιστοιχούν στο πόσο προβλεπτικό είναι κάθε χαρακτηριστικό για κάθε κατηγορία - οι τιμές των βαρών μαθαίνονται με κατάβαση πλαγιάς.
- Για την επισημείωση POS θα κατασκευάζαμε 45 διαφορετικά μοντέλα λογιστικής παλινδρόμησης, ένα για κάθε μέρος του λόγου

Λογιστική παλινδρόμηση (2/4)

- Τα χαρακτηριστικά μπορεί να εξαρτώνται από την ακριβή ταυτότητα μιας λέξης, από κάποια στοιχεία της ορθογραφίας της, ή από κάποιον προσδιορισμό της στην καταχώριση ενός λεξικού.
- Παραδείγματα χαρακτηριστικών:

$w_{i-1} =$ «I»	$w_{i+1} =$ «for»
$w_{i-1} =$ «you»	$c_{i-1} =$ IN
w_i τελειώνει σε «ous»	w_i περιέχει ενωτικό
w_i τελειώνει σε «ly»	w_i περιέχει ψηφίο
w_i αρχίζει με «un»	w_i περιέχει μόνο κεφαλαία
$w_{i-2} =$ «to» και $c_{i-1} =$ VB	w_{i-2} έχει χρονικό προσδιορισμό για το PRESENT
$w_{i-1} =$ «I» και $w_{i+1} =$ «to»	w_{i-2} έχει χρονικό προσδιορισμό για το PAST

Συνολικά μπορεί να υπάρχουν ένα εκατομμύριο χαρακτηριστικά, αλλά για κάθε δεδομένη λέξη μόνο μερικές δεκάδες θα είναι μη μηδενικά.

Λογιστική παλινδρόμηση (3/4)

- Η λογιστική παλινδρόμηση δεν έχει την έννοια της ακολουθίας εισόδων –της παρέχουμε ένα μόνο διάνυσμα χαρακτηριστικών (πληροφορίες για μία μόνο λέξη) και παράγει μια έξοδο (ετικέτα).
- Μπορούμε να αναγκάσουμε τη λογιστική παλινδρόμηση να χειριστεί μια ακολουθία με **άπληστη αναζήτηση** (greedy search): ξεκινάμε επιλέγοντας την πιο πιθανή κατηγορία για την πρώτη λέξη, και συνεχίζουμε με τις υπόλοιπες λέξεις από αριστερά προς τα δεξιά. Σε κάθε βήμα, η κατηγορία *ci* αντιστοιχίζεται σύμφωνα με τον τύπο:

$$c_i = \underset{c' \in Κατηγορία}{\operatorname{argmax}} P(c' \mid w_{1:N}, c_{1:i-1})$$

- Δηλαδή, ο ταξινομητής μπορεί να λάβει υπόψη και τις κατηγορίες προηγούμενων λέξεων που έχουν ήδη αντιστοιχιστεί.

Λογιστική παλινδρόμηση (4/4)

- Μια εναλλακτική προσέγγιση, ενδιάμεση της άπληστης αναζήτησης και του αλγόριθμου Viterbi, είναι η **ακτινική αναζήτηση** (beam search).
- Στην ακτινική αναζήτηση εξετάζουμε κάθε δυνατή κατηγορία σε κάθε χρονικό βήμα, αλλά στη συνέχεια κρατάμε μόνο τις b πιο πιθανές ετικέτες, απορρίπτοντας τις άλλες λιγότερο πιθανές. –Αλλάζοντας την τιμή του b αντισταθμίζουμε μεταξύ ταχύτητας και ακρίβειας.
- Τα απλοϊκά μοντέλα Bayes και τα κρυφά μοντέλα Markov είναι **παραγωγικά μοντέλα** (generative models), ενώ η λογιστική παλινδρόμηση είναι **διαχωριστικό μοντέλο** (discriminative model), με αποτέλεσμα να μην μπορεί να παράγει τυχαίες προτάσεις.
- Συνήθως τα διαχωριστικά μοντέλα έχουν χαμηλότερο ποσοστό σφαλμάτων, αλλά τα παραγωγικά μοντέλα κατασκευάζονται γρηγορότερα.

Σύγκριση γλωσσικών μοντέλων

- Παρακάτω φαίνονται διάφορες προτάσεις που παρήχθησαν τυχαία από μονογραμμικά, διγραμμικά, τριγραμμικά και τετραγραμμικά μοντέλα (κατασκευασμένα από τις λέξεις του αγγλικού πρωτοτύπου αυτού του βιβλίου):

$n = 1$ *logical are as are confusion a may right tries agent goal the was*
$n = 2$: *systems are very similar computational approach would be represented*
$n = 3$: *planning and scheduling are integrated the success of naive Bayes model is*
$n = 4$: *taking advantage of the structure of Bayesian networks and developed various languages for writing "templates" with logical variables, from which large networks could be constructed automatically for each problem instance*

- Το παραπάνω αποτέλεσμα ωστόσο εξαρτάται από το σώμα κειμένων που χρησιμοποιήθηκε για την κατασκευή των μοντέλων.
- Καθώς αυξάνεται το *n*, αυξάνεται και η ποιότητα της γλώσσας που παράγουν, αλλά αντί να παράγουν νέο κείμενο, συνήθως αναπαράγουν αυτολεξεί μεγάλα αποσπάσματα από τα δεδομένα εκπαίδευσής τους

Γραμματική

Εισαγωγή

- **Γραμματική** (grammar) ονομάζεται ένα σύνολο κανόνων που καθορίζουν τη δομή δένδρου των επιτρεπόμενων φράσεων
- **Γλώσσα** (language) είναι το σύνολο των προτάσεων που ακολουθούν αυτούς τους κανόνες.
- Οι **συντακτικές κατηγορίες** (syntactic categories) όπως οι ονοματικές φράσεις (noun phrase) ή ρηματικές φράσεις (verb phrase) μας βοηθούν να περιορίζουμε τις πιθανές λέξεις σε κάθε σημείο μιας πρότασης
- Η **δομή των φράσεων** (phrase structure) παρέχει ένα πλαίσιο για τη σημασία ή τη **σημασιολογία** (semantics) της πρότασης.

Πιθανοτική γραμματική ανεξάρτητη από τα συμφραζόμενα

- Μία **πιθανοτική γραμματική ανεξάρτητη από τα συμφραζόμενα** (probabilistic context-free grammar, PCFG) αποδίδει μια πιθανότητα σε κάθε συμβολοσειρά.
  - Η ιδιότητά της να είναι ανεξάρτητη από τα συμφραζόμενα σημαίνει ότι οποιοσδήποτε κανόνας μπορεί να χρησιμοποιηθεί σε οποιοδήποτε συμφραστικό πλαίσιο: αν η ίδια φράση εμφανίζεται σε δύο θέσεις, πρέπει να έχει κάθε φορά την ίδια πιθανότητα.
- Θα ορίσουμε μια γραμματική PCFG για ένα πολύ μικρό τμήμα αγγλικού κειμένου το οποίο αφορά την επικοινωνία μεταξύ πρακτόρων που εξερευνούν τον κόσμο του wumpus (wumpus world).
  - Ονομάζουμε αυτήν τη γλώσσα $\mathcal{E}0$ (Εικόνα 23.2).
- Δυστυχώς, η γραμματική $\mathcal{E}0$ **υπερπαράγει** (overgenerates) αλλά και **υποπαράγει** (undergenerates) προτάσεις.

Π	→	*ΟΦ ΡΦ*	[0,90]	I + feel a breeze
	|	*Π Σύνδεσμος Π*	[0,10]	I feel a breeze + and + It stinks
ΟΦ	→	*Αντωνυμία*	[0,25]	I
	|	*Όνομα*	[0,10]	Ali
	|	*Ουσιαστικό*	[0,10]	pits
	|	*Άρθρο Ουσιαστικό*	[0,25]	the + wumpus
	|	*Άρθρο ΕΠ Ουσιαστικό*	[0,05]	the + smelly dead + wumpus
	|	*Ψηφίο Ψηφίο*	[0,05]	3 4
	|	*ΟΦ ΠΦ*	[0,10]	the wumpus + in 1 3
	|	*ΟΦ ΑναφΠροτ*	[0,05]	the wumpus + that is smelly
	|	*ΟΦ Σύνδεσμος ΟΦ*	[0,05]	the wumpus + and + I
ΡΦ	→	*Ρήμα*	[0,40]	stinks
	|	*ΡΦ ΟΦ*	[0,35]	feel + a breeze
	|	*ΡΦ Επίθετο*	[0,05]	smells + dead
	|	*ΡΦ ΠΦ*	[0,10]	is + in 1 3
	|	*ΡΦ Επίρρημα*	[0,10]	go + ahead
ΕΠ	→	*Επίθετο*	[0,80]	smelly
	|	*Επίθετο ΕΠ*	[0,20]	smelly + dead
ΠΦ	→	*Πρόθεση ΟΦ*	[1,00]	to + the east
ΑναφΠροτ	→	*ΑναφΑντ ΡΦ*	[1,00]	that + is smelly

Εικόνα 23.2 Η γραμματική της $\mathscr{E}_0$, με παραδείγματα φράσεων για κάθε κανόνα. Οι συντακτικές κατηγορίες είναι οι εξής: πρόταση (*Π*), ονοματική φράση (*ΟΦ*), ρηματική φράση (*ΡΦ*), λίστα επιθέτων (*ΕΠ*), προθετική φράση (*ΠΦ*), και αναφορική πρόταση (*ΑναφΠροτ*).

Το λεξικό της ε_0

- Μέρος του **λεξικού** (lexicon), δηλαδή της λίστας των επιτρεπόμενων λέξεων, ορίζεται στην Εικόνα 23.3.
  - Τα ουσιαστικά, τα κύρια ονόματα, τα ρήματα, τα επίθετα και τα επιρρήματα, είναι **ανοιχτές κατηγορίες** (open classes).
  - Οι αντωνυμίες, οι αναφορικές αντωνυμίες, τα άρθρα, οι προθέσεις και οι σύνδεσμοι είναι **κλειστές κατηγορίες** (closed classes).

Ουσιαστικό	→	**stench** [0,05] | **breeze** [0,10] | **wumpus** [0,15] | **pits** [0,05] | ...
Ρήμα	→	**is** [0,10] | **feel** [0,10] | **smells** [0,10] | **stinks** [0,05] | ...
Επίθετο	→	**right** [0,10] | **dead** [0,05] | **smelly** [0,02] | **breezy** [0,02] ...
Επίρρημα	→	**here** [0,05] | **ahead** [0,05] | **nearby** [0,02] | ...
Αντωνυμία	→	**me** [0,10] | **you** [0,03] | **I** [0,10] | **it** [0,10] | ...
ΑναφΑντ	→	**that** [0,40] | **which** [0,15] | **who** [0,20] | **whom** [0,02] | ...
Όνομα	→	**Ali** [0,01] | **Bo** [0,01] | **Boston** [0,01] | ...
Άρθρο	→	**the** [0,40] | **a** [0,30] | **an** [0,10] | **every** [0,05] | ...
Πρόθεση	→	**to** [0,20] | **in** [0,10] | **on** [0,05] | **near** [0,10] | ...
Σύνδεσμος	→	**and** [0,50] | **or** [0,10] | **but** [0,20] | **yet** [0,02] | ...
Ψηφίο	→	**0** [0,20] | **1** [0,20] | **2** [0,20] | **3** [0,20] | **4** [0,20] | ...

Εικόνα 23.3 Το λεξικό της $\mathscr{E}_0$. Το *ΑναφΑντ* είναι συντομογραφία για τις αναφορικές αντωνυμίες. Το άθροισμα των πιθανοτήτων για κάθε κατηγορία είναι 1.

Συντακτική Ανάλυση

Εισαγωγή

- **Συντακτική ανάλυση** (parsing) είναι η διαδικασία ανάλυσης μιας συμβολοσειράς λέξεων με σκοπό την αποκάλυψη της δομής των φράσεών της σύμφωνα με τους κανόνες μιας γραμματικής.
  - Μπορούμε να τη θεωρούμε ως την **αναζήτηση** (search) ενός έγκυρου συντακτικού δένδρου του οποίου τα φύλλα είναι οι λέξεις της συμβολοσειράς.
- Εικόνα 23.4.
- Μπορούμε να ξεκινήσουμε με το σύμβολο Π και να εκτελέσουμε την αναζήτηση από επάνω προς τα κάτω, ή να ξεκινήσουμε με τις λέξεις και να εκτελέσουμε την αναζήτηση από κάτω προς τα επάνω.
  - Ωστόσο, οι αμιγείς στρατηγικές συντακτικής ανάλυσης από επάνω προς τα κάτω ή από κάτω προς τα επάνω μπορεί να μην είναι αποδοτικές.

Λίστα στοιχείων	Κανόνας
Π	
ΟΦ ΡΦ	*Π* $\rightarrow$ *ΟΦ ΡΦ*
ΟΦ ΡΦ Επίθετο	*ΡΦ* $\rightarrow$ *ΡΦ Επίθετο*
ΟΦ Ρήμα Επίθετο	*ΡΦ* $\rightarrow$ *Ρήμα*
ΟΦ Ρήμα **dead**	*Επίθετο* $\rightarrow$ **dead**
ΟΦ **is dead**	*Ρήμα* $\rightarrow$ **is**
Άρθρο Ουσιαστικό **is dead**	*ΟΦ* $\rightarrow$ *Άρθρο Ουσιαστικό*
Άρθρο **wumpus is dead**	*Ουσιαστικό* $\rightarrow$ **wumpus**
the wumpus is dead	*Άρθρο* $\rightarrow$ **the**

Εικόνα 23.4 Συντακτική ανάλυση της συμβολοσειράς «The wumpus is dead» ως πρότασης, σύμφωνα με τη γραμματική της $\mathscr{E}_0$. Αν εκτελέσουμε συντακτική ανάλυση από επάνω προς τα κάτω, ξεκινάμε με το *Π*, και σε κάθε βήμα αντιστοιχίζουμε ένα μη τερματικό X με έναν κανόνα της μορφής ($X \rightarrow Y\ldots$) και αντικαθιστούμε το X με το $Y\ldots$ στη λίστα των στοιχείων· για παράδειγμα, αντικαθιστούμε το *Π* με την ακολουθία *ΟΦ ΡΦ*. Αν εκτελέσουμε συντακτική ανάλυση από κάτω προς τα επάνω, ξεκινάμε με τις λέξεις «the wumpus is dead», και σε κάθε βήμα αντιστοιχίζουμε μια συμβολοσειρά λεκτικών μονάδων όπως την ($Y\ldots$) με έναν κανόνα της μορφής ($X \rightarrow Y\ldots$) και αντικαθιστούμε τις λεκτικές μονάδες με το X· για παράδειγμα, αντικαθιστούμε το «the» με *Άρθρο* ή το *Άρθρο Ουσιαστικό* με *ΟΦ*.

Δυναμικός προγραμματισμός

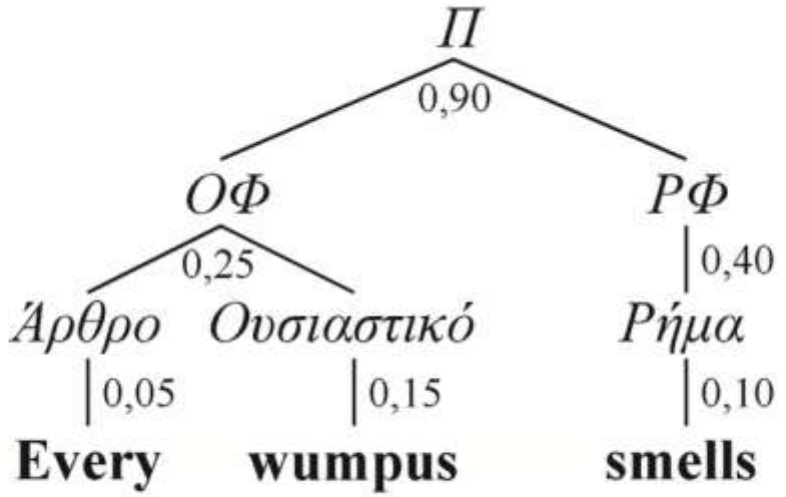
- Καλύτερα αποτελέσματα επιτυγχάνονται με **δυναμικό προγραμματισμό** (dynamic programming)
  - Κάθε φορά που αναλύουμε μια υποσυμβολοσειρά, αποθηκεύουμε τα αποτελέσματα ώστε να μην χρειάζεται να επαναλάβουμε την ίδια ανάλυση αργότερα.
- Τα ενδιάμεσα αποτελέσματα καταγράφονται σε μια δομή δεδομένων γνωστή ως **διάγραμμα** (chart). Ένας αλγόριθμος που το κάνει αυτό ονομάζεται **διαγραμματικός συντακτικός αναλυτής** (chart parser).
- Υπάρχουν πολλοί τύποι διαγραμματικών συντακτικών αναλυτών.
- Στο βιβλίο των Norvig και Russell (εικόνα 23.5) παρουσιάζεται η πιθανοτική εκδοχή ενός αλγορίθμου συντακτικής ανάλυσης από κάτω προς τα επάνω που ονομάζεται CYK, από τα αρχικά των δημιουργών του, Ali Cocke, Daniel Younger και Tadeo Kasami.

Ο αλγόριθμος CYK

- Ο αλγόριθμος CYK απαιτεί μια γραμματική με όλους τους κανόνες σε μία από τις δύο πολύ συγκεκριμένες μορφές:
  - λεξικολογικούς κανόνες της μορφής X → λέξη [p]
  - συντακτικούς κανόνες της μορφής X→ YZ[p], με ακριβώς δύο κατηγορίες στο δεξιό μέρος τους.
- Αυτή η μορφή γραμματικής, που ονομάζεται κανονική μορφή Chomsky (Chomsky Normal Form).
  - Οποιαδήποτε γραμματική που είναι ανεξάρτητη από τα συμφραζόμενα μπορεί να μετασχηματιστεί αυτόματα σε κανονική μορφή Chomsky.
- Ο αλγόριθμος CYK χρησιμοποιεί χώρο O(n2m) για τους πίνακες P και T, όπου n είναι ο αριθμός των λέξεων στην πρόταση και m ο αριθμός των μη τερματικών συμβόλων στη γραμματική, και απαιτεί χρόνο O(n3m)

Συντακτική ανάλυση με αναζήτηση A*

- Για να προσπαθήσουμε να φτάσουμε σε χρόνο O(n) μπορούμε να εφαρμόσουμε αναζήτηση A* με έναν σχετικά απλό τρόπο:
  - Κάθε κατάσταση είναι μια λίστα στοιχείων (λέξεων ή κατηγοριών), όπως φαίνεται στην Εικόνα 23.4.
  - Η αρχική κατάσταση είναι μια λίστα λέξεων, και μια κατάσταση στόχου είναι το μεμονωμένο στοιχείο Π.
  - Το κόστος μιας κατάστασης είναι το αντίστροφο της πιθανότητάς της όπως ορίζεται από τους κανόνες που έχουμε εφαρμόσει μέχρι τώρα, και υπάρχουν διάφορες ευρετικές μέθοδοι για την εκτίμηση της εναπομένουσας απόστασης από τον στόχο.
- Οι καλύτερες ευρετικές μέθοδοι που χρησιμοποιούνται σήμερα προέρχονται από μηχανική μάθηση που εφαρμόζεται σε ένα σώμα προτάσεων.
- Ένα παράδειγμα συντακτικής ανάλυσης με A* παρουσιάζεται στην Εικόνα 23.6.



Εικόνα 23.6 Συντακτικό δένδρο για την πρόταση «Every wumpus smells», σύμφωνα με τη γραμματική της $\mathscr{E}_0$. Σε κάθε εσωτερικό κόμβο του δένδρου αναγράφεται η πιθανότητά του. Η πιθανότητα του δένδρου στο σύνολό του είναι $0,9 \times 0,25 \times 0,05 \times 0,15 \times 0,40 \times 0,10 = 0,0000675$. Το δένδρο μπορεί επίσης να γραφεί σε γραμμική μορφή ως [*Π* [*ΟΦ* [*Άρθρο* **every**] [*Ουσιαστικό* **wumpus**]][*ΡΦ* [*Ρήμα* **smells**]]].

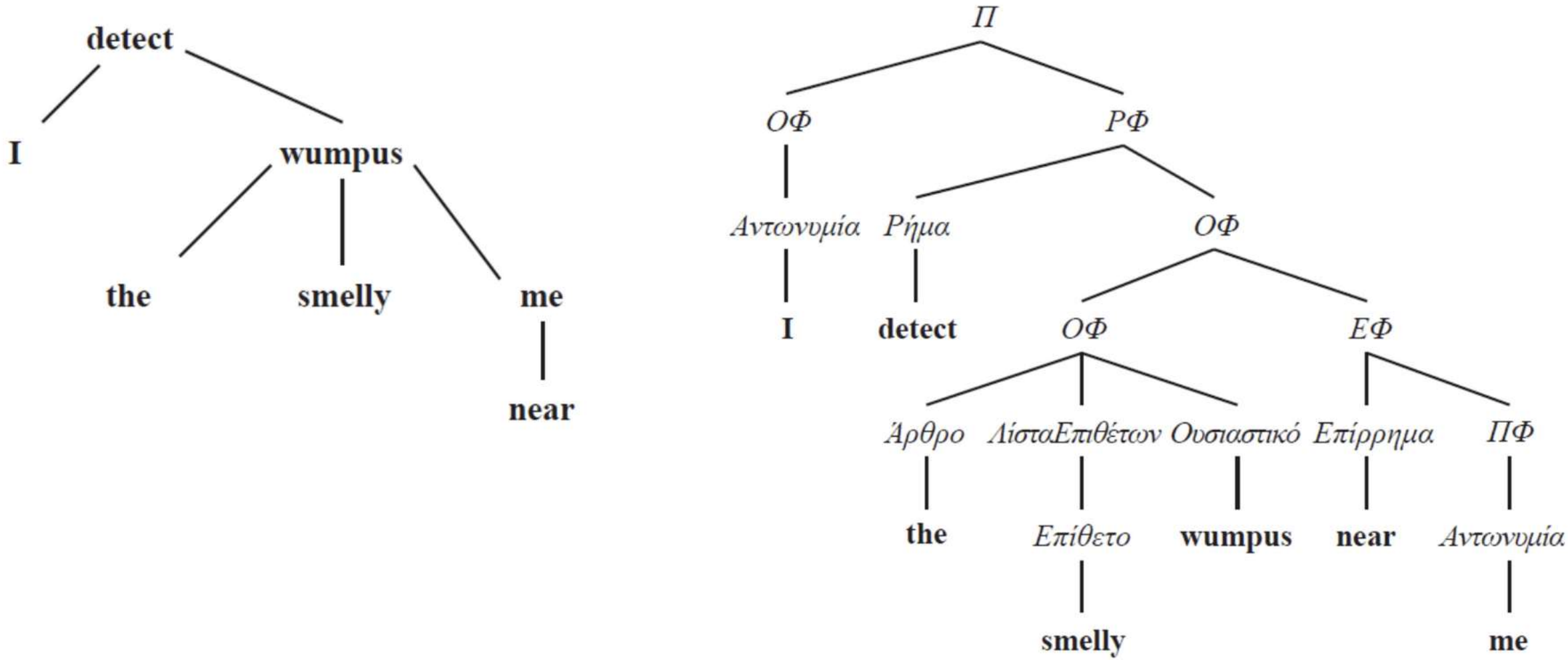
- Η πρώτη συντακτική ανάλυση που θα βρεθεί θα είναι η πιο πιθανή, με την προϋπόθεση μιας παραδεκτής ευρετικής μεθόδου

Συντακτική ανάλυση με ακτινική αναζήτηση

- Μπορούμε να χρησιμοποιήσουμε **ακτινική αναζήτηση** (beam search) για συντακτική ανάλυση, όπου οποιαδήποτε στιγμή θα εξετάζουμε μόνο τις b πιο πιθανές εναλλακτικές συντακτικές αναλύσεις.
- Ένας συντακτικός αναλυτής ακτινικής αναζήτησης με b=1ονομάζεται **αιτιοκρατικός συντακτικός αναλυτής** (deterministic parser).
- Μια δημοφιλής αιτιοκρατική προσέγγιση είναι η **συντακτική ανάλυση ολίσθησης-ελάττωσης** (shift-reduce parsing), στην οποία εξετάζουμε την πρόταση λέξη προς λέξη, και σε κάθε σημείο επιλέγουμε αν θα μεταφέρουμε τη λέξη σε μια στοίβα συστατικών μερών (ολίσθηση) ή αν θα αφαιρέσουμε τα συστατικά στην κορυφή της στοίβας σύμφωνα με κάποιον γραμματικό κανόνα (ελάττωση)

Συντακτική ανάλυση βάσει εξαρτήσεων

- Υπάρχει μια ευρέως χρησιμοποιούμενη εναλλακτική συντακτική προσέγγιση που ονομάζεται **γραμματική εξαρτήσεων** (dependency grammar) και υποθέτει ότι η συντακτική δομή σχηματίζεται από δυαδικές σχέσεις μεταξύ λεξικολογικών στοιχείων, χωρίς να απαιτούνται συντακτικά συστατικά μέρη.
  - Η Εικόνα 23.7 παρουσιάζει τη συντακτική ανάλυση των εξαρτήσεων και της δομής των φράσεων μιας πρότασης.
- Η δημοτικότητα των γραμματικών εξαρτήσεων οφείλεται σε μεγάλο βαθμό στο έργο Universal Dependencies (Nivre κ.ά., 2016), ένα έργο treebank (τράπεζα δένδρων) ανοιχτού πηγαίου κώδικα που ορίζει ένα σύνολο σχέσεων και παρέχει εκατομμύρια συντακτικά αναλυμένες προτάσεις σε περισσότερες από 70 γλώσσες.



Εικόνα 23.7 Συντακτική ανάλυση βάσει εξαρτήσεων (επάνω) και η αντίστοιχη συντακτική ανάλυση της δομής των φράσεων (κάτω) για την πρόταση *I detect the smelly wumpus near me.*

Εκπαίδευση συντακτικού αναλυτή από παραδείγματα

- Θα ήταν καλύτερο να μαθαίνουμε τους γραμματικούς κανόνες (και τις πιθανότητές τους) αντί να τους γράφουμε με το χέρι.
- Με δεδομένο ένα treebank, μπορούμε να δημιουργήσουμε μια γραμματική PCFG μετρώντας απλώς τον αριθμό των φορών που κάθε τύπος κόμβου εμφανίζεται σε ένα δένδρο (με τους συνηθισμένους περιορισμούς σχετικά με την εξομάλυνση των χαμηλών μετρήσεων).
- Στην Εικόνα 23.8, υπάρχουν δύο κόμβοι της μορφής [S[NP. . .][VP. . .]].
- Θα μετρούσαμε αυτούς και όλα τα άλλα υποδένδρα με ρίζα Sστο σώμα κειμένων.
- Αν υπάρχουν 1000 κόμβοι S από τους οποίους οι 600 είναι αυτής της μορφής, τότε δημιουργούμε τον κανόνα:
  - –S → NP VP [0,6]

```
[ [S [NP-2 Her eyes]
     [VP were
          [VP glazed
               [NP *-2]
               [SBAR-ADV as if
                    [S [NP she]
                         [VP did n't
                              [VP [VP hear [NP *-1]]
                                   or
                                   [VP [ADVP even] see [NP *-1]]
                                   [NP-1 him]]]]]]]
  .]
```

Εικόνα 23.8 Επισημειωμένο δένδρο για την πρόταση «Her eyes were glazed as if she didn't hear or even see him.» από το σώμα κειμένων Penn Treebank. Προσέξτε ένα γραμματικό φαινόμενο που δεν έχουμε καλύψει ακόμα: τη μετακίνηση μιας φράσης από ένα τμήμα του δένδρου σε ένα άλλο. Αυτό το δένδρο αναλύει τη φράση «hear or even see him» ως αποτελούμενη από δύο ρηματικές φράσεις, [*VP* **hear** [*NP* *-1]] και [*VP* [*ADVP* **even**] **see** [*NP* *-1]]· και από τις δύο λείπει το αντικείμενο που επισημειώνεται ως *-1 και αφορά το στοιχείο *NP* που σε όλο το υπόλοιπο δένδρο χαρακτηρίζεται ως [*NP*-1 **him**]. Παρόμοια, το [*NP* *-2] αναφέρεται στο [*NP*-2 **Her eyes**].

Συντακτική ανάλυση προσανατολισμένη στα δεδομένα

- Έστω η φράση «the good and the bad».
- Η φράση αυτή αναλύεται καταρχήν ως μία μοναδική ονοματική φράση και όχι ως δύο συνδεόμενες ονοματικές φράσεις, γεγονός που μας δίνει τον κανόνα:
  - NP → Άρθρο Ουσιαστικό Σύνδεσμος Άρθρο Ουσιαστικό
- Μια πιο περιεκτική γραμματική θα μπορούσε να αποτυπώσει όλους τους κανόνες συνδεόμενων ονοματικών φράσεων σε έναν και μοναδικό κανόνα:
  - NP → NP Σύνδεσμος NP
- Οι Bod κ.ά. (2003) και ο Bod (2008) δείχνουν πώς μπορούμε να δημιουργούμε αυτόματα γενικευμένους κανόνες σαν και αυτόν, μειώνοντας σημαντικά τον αριθμό των κανόνων που προκύπτουν από το treebank, και δημιουργώντας μια γραμματική η οποία τελικά θα γενικεύει καλύτερα για προτάσεις που δεν είχαν εξεταστεί προηγουμένως.
  - Οι ίδιοι αποκαλούν την προσέγγισή τους **συντακτική ανάλυση προσανατολισμένη στα δεδομένα** (data-oriented parsing).

Μη επιβλεπόμενη συντακτική ανάλυση

- Μια εναλλακτική προσέγγιση είναι η **μη επιβλεπόμενη συντακτική ανάλυση** (unsupervised parsing), στην οποία μαθαίνουμε μια νέα γραμματική (ή βελτιώνουμε κάποια υπάρχουσα) χρησιμοποιώντας ένα σώμα προτάσεων χωρίς δένδρα.
- Ο **αλγόριθμος μέσα–έξω** (inside–outside algorithm) (Dodd, 1988), τον οποίο δεν θα περιγράψουμε εδώ, μαθαίνει να εκτιμά τις πιθανότητες σε μια γραμματική PCFG μέσα από παραδείγματα προτάσεων χωρίς δένδρα, με τρόπο παρόμοιο με τον αλγόριθμο εμπρός-πίσω (Εικόνα 14.4 –forward-backward algorithm).
- Οι Spitkovsky κ.ά. (2010a) περιγράφουν μια προσέγγιση μη επιβλεπόμενης μάθησης που χρησιμοποιεί **μάθηση με πρόγραμμα** (curriculum learning): ξεκινάμε με το εύκολο κομμάτι του προγράμματος και σταδιακά μαθαίνουμε μεγαλύτερες και προτάσεις.

Ημι-επιβλεπόμενη συντακτική ανάλυση

- Στην **ημι-επιβλεπόμενη συντακτική ανάλυση** (semisupervised parsing) ξεκινάμε με ένα μικρό πλήθος δένδρων τα οποία χρησιμοποιούμε ως δεδομένα για τη δημιουργία μιας αρχικής γραμματικής, και στη συνέχεια προσθέτουμε ένα μεγάλο πλήθος μη αναλυμένων προτάσεων για τη βελτίωσή της.
- Μπορούμε να χρησιμοποιήσουμε **μερική επισημείωση** (partial bracketing): ευρέως διαθέσιμα κείμενα που έχουν επισημειωθεί από τους συγγραφείς, και όχι από ειδικούς γλωσσολόγους, με μια μερική δενδροειδή δομή, σε μορφή HTML ή παρόμοιων επισημειώσεων.
  - Σε κείμενα HTML οι περισσότερες αγκύλες αντιστοιχούν σε κάποιο συντακτικό στοιχείο, οπότε η μερική επισημείωση μπορεί να βοηθήσει στη μάθηση μιας γραμματικής.

ΒΑΘΙΑ ΜΑΘΗΣΗ ΓΙΑ ΕΠΕΞΕΡΓΑΣΙΑ ΦΥΣΙΚΗΣ ΓΛΩΣΣΑΣ

Ενσωματώσεις λέξεων

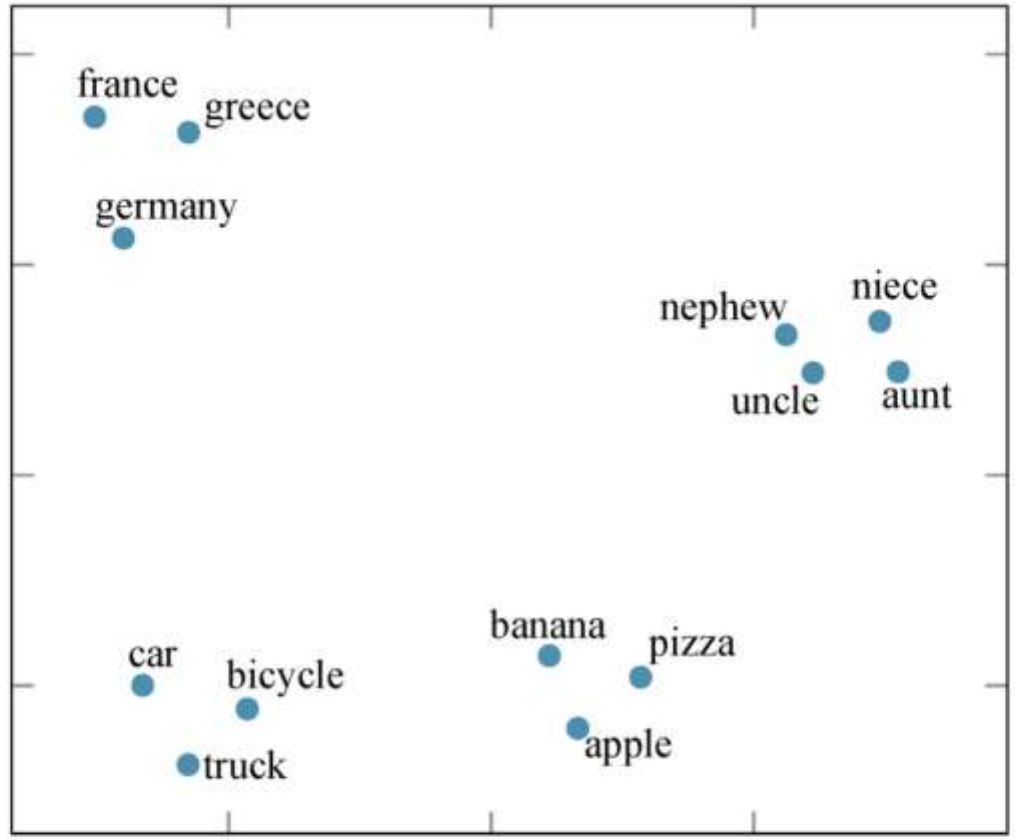
Εισαγωγή

- Επιθυμούμε μια αναπαράσταση λέξεων που θα επιτρέπει τη γενίκευση μεταξύ σχετικών λέξεων –λέξεων που σχετίζονται:
  - συντακτικά (τα «άχρωμος» και «ιδανικός» είναι και τα δύο επίθετα)
  - σημασιολογικά (τα «γάτα» και «γατάκι» είναι και τα δύο αιλουροειδή)
  - θεματικά (τα «ηλιόλουστος» και «χιονόνερο» είναι και τα δύο μετεωρολογικοί όροι
  - συναισθηματικά (το «φανταστικά» προκαλεί ακριβώς το αντίθετο συναίσθημα από το «απαίσια»),
  - ή με κάποιον άλλον τρόπο.
- Η αναπαράσταση λέξεων με διανύσματα **one-hot** δεν αποτυπώνει την ομοιότητα μεταξύ των λέξεων.

Ενσωμάτωση λέξης

- Σύμφωνα με τη ρήση του γλωσσολόγου JohnR. Firth(1957), «Γνωρίζουμε μια λέξη από αυτές που τη συνοδεύουν».
- Μια **ενσωμάτωση λέξης** (word embedding) είναι ένα πυκνό διάνυσμα πραγματικών αριθμών το οποίο αναπαριστά μια λέξη.
  - Σε αντιδιαστολή, το διάνυσμα one-hot έχει πάρα πολλές διαστάσεις (αραιό διάνυσμα) με τιμές 0 και 1.
- Οι ενσωματώσεις λέξεων μαθαίνονται αυτόματα από τα δεδομένα.
  - Μπορεί κανείς να κατασκευάσει διαφορετικά σύνολα ενσωματώσεων για τις ίδιες λέξεις και από τα ίδια δεδομένα.
- Παραδείγματα ενσωματώσεων:

$$
\begin{aligned}
\text{«aardvark»} &= [-0{,}7, +0{,}2, -3{,}2, \ldots] \\
\text{«abacus»} &= [+0{,}5, +0{,}9, -1{,}3, \ldots] \\
&\ldots \\
\text{«zyzzyva»} &= [-0{,}1, +0{,}8, -0{,}4, \ldots]
\end{aligned}
$$



Εικόνα 24.1 Διανύσματα ενσωματώσεων λέξεων, υπολογισμένα από τον αλγόριθμο GloVe που έχει εκπαιδευτεί σε 6 δισεκατομμύρια λέξεις κειμένου. Σε αυτήν την απεικόνιση, διανύσματα λέξεων 100 διαστάσεων προβάλλονται σε δύο διαστάσεις. Οι παρόμοιες λέξεις εμφανίζονται κοντά η μία στην άλλη.

Ιδιότητες ενσωματώσεων

- Αποδεικνύεται (πειραματικά), για λόγους που δεν κατανοούμε πλήρως, ότι τα διανύσματα ενσωμάτωσης λέξεων έχουν πρόσθετες ιδιότητες πέραν της απλής εγγύτητάς τους για παρόμοιες λέξεις.
  - Για παράδειγμα, έστω τα διανύσματα **A** για τη λέξη «Αθήνα» και **B** για τη λέξη «Ελλάδα».
  - Για αυτές τις λέξεις, η διαφορά των διανυσμάτων τους, **B−A**, φαίνεται να κωδικοποιεί τη σχέση χώρας και πρωτεύουσας.
  - Άλλα ζεύγη, π.χ., Γαλλία και Παρίσι, Ρωσία και Μόσχα, Ζάμπια και Λουσάκα, έχουν ουσιαστικά την ίδια διαφορά διανυσμάτων.
- Μπορούμε να χρησιμοποιούμε αυτήν την ιδιότητα για να επιλύουμε προβλήματα αναλογίας λέξεων, όπως «Η Αθήνα (**A**) είναι για την Ελλάδα (**B**) ό,τι το Όσλο (**C**)για [τι] (**D**);».
- **– D=C+(B−A)**

A	**B**	**C**	**D = C + (B − A)**	**Σχέση**
Athens	Greece	Oslo	Norway	*Capital*
Astana	Kazakhstan	Harare	Zimbabwe	*Capital*
Angola	kwanza	Iran	rial	*Currency*
copper	Cu	gold	Au	*Atomic Symbol*
Microsoft	Windows	Google	Android	*Operating System*
New York	New York Times	Baltimore	Baltimore Sun	*Newspaper*
Berlusconi	Silvio	Obama	Barack	*First name*
Switzerland	Swiss	Cambodia	Cambodian	*Nationality*
Einstein	scientist	Picasso	painter	*Occupation*
brother	sister	grandson	granddaughter	*Family Relation*
Chicago	Illinois	Stockton	California	*State*
possibly	impossibly	ethical	unethical	*Negative*
mouse	mice	dollar	dollars	*Plural*
easy	easiest	lucky	luckiest	*Superlative*
walking	walked	swimming	swam	*Past tense*

Εικόνα 24.2 Ένα μοντέλο ενσωματώσεων λέξεων μπορεί μερικές φορές να απαντά στην ερώτηση «Το **A** είναι για το **B** ό,τι το **C** για [τι];» με αριθμητική διανυσμάτων: Με δεδομένα τα διανύσματα ενσωματώσεων λέξεων για τις λέξεις **A**, **B** και **C**, υπολογίζουμε το διάνυσμα $\mathbf{D} = \mathbf{C} + (\mathbf{B} - \mathbf{A})$ και αναζητούμε τη λέξη που είναι πιο κοντά στο **D**. (Οι απαντήσεις στη στήλη **D** έχουν υπολογιστεί αυτόματα από το μοντέλο. Οι περιγραφές στη στήλη «Σχέση» έχουν προστεθεί με το χέρι.) Κατόπιν προσαρμογής από τον Mikolov κ.ά. (2013, 2014).

Προεκπαιδευμένα διανύσματα

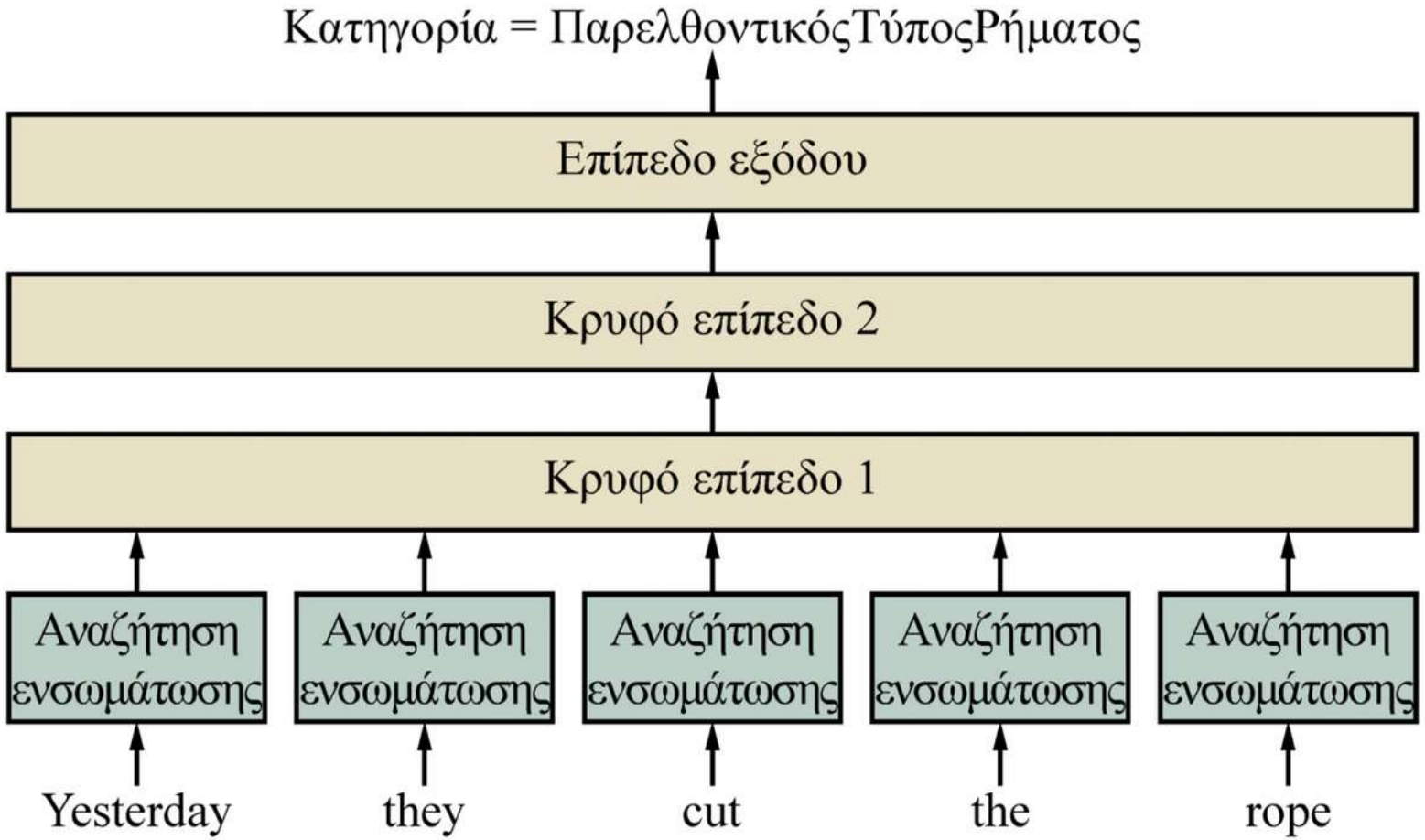
- Δεν υπάρχει ωστόσο καμία εγγύηση ότι ένας συγκεκριμένος αλγόριθμος ενσωμάτωσης λέξεων που εκτελείται σε ένα συγκεκριμένο σώμα κειμένων θα αποτυπώσει μια συγκεκριμένη σημασιολογική σχέση.
- Η χρήση διανυσμάτων ενσωματώσεων λέξεων αντί κωδικοποιήσεων λέξεων one-hot αποδεικνύεται χρήσιμη για όλες σχεδόν τις εφαρμογές βαθιάς μάθησης σε εργασίες NLP.
  - Σε πολλές περιπτώσεις συγκεκριμένων εργασιών NLP είναι δυνατό να χρησιμοποιούνται γενικά προεκπαιδευμένα (pretrained) διανύσματα, τα οποία λαμβάνονται από οποιαδήποτε από τις διάφορες διαθέσιμες πηγές.
- Στα πιο συνηθισμένα λεξικά διανυσμάτων περιλαμβάνονται τα WORD2VEC, GLOVE (Global Vectors), και FASTTEXT, εκ των οποίων το τελευταίο περιέχει ενσωματώσεις για 157 γλώσσες.

Εκπαίδευση για συγκεκριμένο πεδίο

- Μπορούμε να εκπαιδεύσουμε δικά μας διανύσματα λέξεων, παράλληλα με την εκπαίδευση ενός δικτύου για μια συγκεκριμένη εργασία.
- Οι ενσωματώσεις αυτές τείνουν να δίνουν έμφαση σε πτυχές των λέξεων που είναι χρήσιμες για αυτήν την εργασία.
- Έστω το πρόβλημα της παραγωγής ενσωματώσεων για την εργασία της επισημείωσης λέξεων (POS tagging). Τα βήματα που πρέπει να ακολουθηθούν είναι τα εξής:
  - Επιλέγουμε το πλάτος w (έναν περιττό αριθμό λέξεων) για το παράθυρο της πρόβλεψης που θα χρησιμοποιηθεί για την επισήμανση κάθε λέξης και παράγουμε επικαλυπτόμενα παραδείγματα από τα δεδομένα.
    - Μια συνηθισμένη τιμή είναι w=5, που σημαίνει ότι η ετικέτα μιας λέξης προβλέπεται με βάση την ίδια τη λέξη συν δύο λέξεις αριστερά και δύο λέξεις δεξιά της.

Εκπαίδευση για συγκεκριμένο πεδίο (2)

- (συνέχεια)
- Δημιουργούμε ένα λεξιλόγιο με όλες τις λεκτικές μονάδες (ή σύμβολα, tokens) που εμφανίζονται, έστω, περισσότερες από 5 φορές στα δεδομένα εκπαίδευσης. Έστω v το μέγεθος του λεξιλογίου.
- Ταξινομούμε αυτό το λεξιλόγιο με οποιαδήποτε τυχαία σειρά (ίσως αλφαβητικά).
- Επιλέγουμε μια τιμή d ως μέγεθος κάθε διανύσματος ενσωμάτωσης μιας λέξης.
- Δημιουργούμε τη μήτρα ενσωματώσεων E διαστάσεων $v \times d$ (word embedding matrix). Αρχικοποιούμε με τυχαίες τιμές.
- Κατασκευάζουμε ένα νευρωνικό δίκτυο που παράγει ετικέτες για τα μέρη του λόγου, όπως φαίνεται στην Εικόνα 24.3. Το πρώτο επίπεδο θα αποτελείται από w αντίγραφα της μήτρας ενσωματώσεων.
- Εκπαιδεύουμε το νευρωνικό δίκτυο.



Οι λέξεις στην είσοδο δίνονται με κωδικοποίηση one-hot.

Εικόνα 24.3 Μοντέλο πρόσθιας τροφοδότησης για επισημείωση POS. Το μοντέλο αυτό λαμβάνει ως είσοδο ένα παράθυρο 5 λέξεων και προβλέπει την ετικέτα της μεσαίας λέξης –εδώ, της *cut*. Μπορεί να λαμβάνει υπόψη τη θέση των λέξεων επειδή κάθε μία από τις 5 ενσωματώσεις εισόδου πολλαπλασιάζεται με ένα διαφορετικό τμήμα του πρώτου κρυφού επιπέδου. Οι τιμές των παραμέτρων για τις ενσωματώσεις λέξεων και για τα τρία επίπεδα μαθαίνονται όλες ταυτόχρονα κατά την εκπαίδευση.

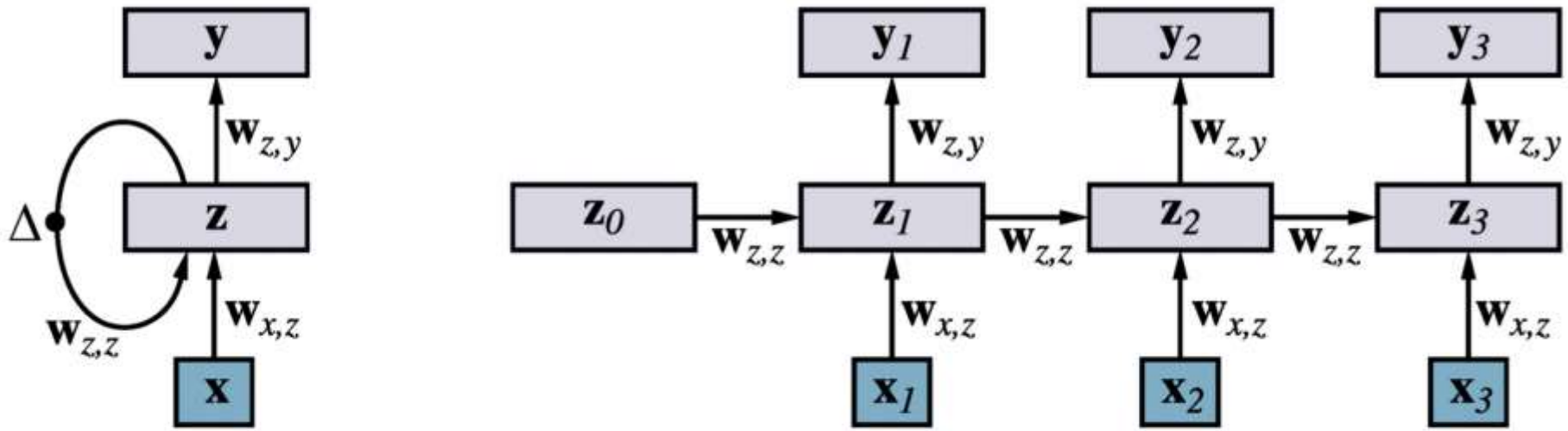
Αναδρομικά νευρωνικά δίκτυα για NLP

Συμφραζόμενα

- Η γλώσσα αποτελείται από διατεταγμένες ακολουθίες λέξεων, στις οποίες τα **συμφραζόμενα** (context) είναι σημαντικά.
- Για απλές εργασίες, όπως η επισημείωση με μέρη του λόγου, ένα μικρό παράθυρο σταθερού μεγέθους, ενδεχομένως πέντε λέξεων, συνήθως παρέχει επαρκή συμφραζόμενα.
- Πιο σύνθετες εργασίες, όπως η απάντηση ερωτήσεων ή η ανάλυση αναφορών, ενδέχεται να απαιτούν ως συμφραζόμενα δεκάδες λέξεις.

Γλωσσικά μοντέλα με αναδρομικά νευρωνικά δίκτυα

- Έστω το πρόβλημα της δημιουργίας ενός γλωσσικού μοντέλου (language model) με επαρκή συμφραζόμενα.
- Η δημιουργία ενός γλωσσικού μοντέλου είτε με ένα n-γραμμικό μοντέλο είτε με ένα δίκτυο πρόσθιας τροφοδότησης με σταθερό παράθυρο n λέξεων μπορεί να αποδειχθεί δύσκολη λόγω του προβλήματος των συμφραζομένων.
  - –Επιπρόσθετα, ένα δίκτυο πρόσθιας τροφοδότησης παρουσιάζει το πρόβλημα της **ασυμμετρίας** (asymmetry): τα βάρη για μια λέξη είναι διαφορετικά για διαφορετικές θέσεις εμφάνισης της λέξης εντός του σταθερού παραθύρου.
- Τα αναδρομικά νευρωνικά δίκτυα (recurrent neural networks, RNN) είναι σχεδιασμένα να επεξεργάζονται δεδομένα χρονοσειρών, ένα προς ένα. Άρα ενδέχεται να είναι χρήσιμα για την επεξεργασία της γλώσσας, λέξη προς λέξη (Εικόνα 24.2).



Εικόνα 24.4 (α) Σχηματικό διάγραμμα δικτύου RNN όπου το κρυφό επίπεδο **z** έχει αναδρομικές συνδέσεις· το σύμβολο Δ υποδηλώνει καθυστέρηση. Κάθε είσοδος **x** είναι το διάνυσμα ενσωμάτωσης της επόμενης λέξης στην πρόταση. Κάθε έξοδος **y** είναι η έξοδος για το συγκεκριμένο χρονικό βήμα. (β) Το ίδιο δίκτυο εκτυλιγμένο σε τρία χρονικά βήματα για τη δημιουργία ενός δικτύου πρόσθιας τροφοδότησης. Σημειώστε ότι τα βάρη είναι κοινά σε όλα τα χρονικά βήματα.

Γλωσσικά μοντέλα με αναδρομικά νευρωνικά δίκτυα

- Σε ένα γλωσσικό μοντέλο RNN, κάθε λέξη εισόδου κωδικοποιείται ως ένα διάνυσμα ενσωματώσεων λέξεων, **xi**. Υπάρχει ένα κρυφό επίπεδο **zt** που μεταβιβάζεται ως είσοδος από το ένα χρονικό βήμα στο επόμενο.
- Μας ενδιαφέρει να πραγματοποιήσουμε ταξινόμηση πολλών κατηγοριών: οι κατηγορίες είναι οι λέξεις του λεξιλογίου.
- Έτσι, η έξοδος **yt** θα είναι μια κατανομή πιθανοτήτων *softmax* για τις δυνατές τιμές της επόμενης λέξης στην πρόταση.

Ιδιότητες γλωσσικών μοντέλων με αναδρομικά νευρωνικά δίκτυα

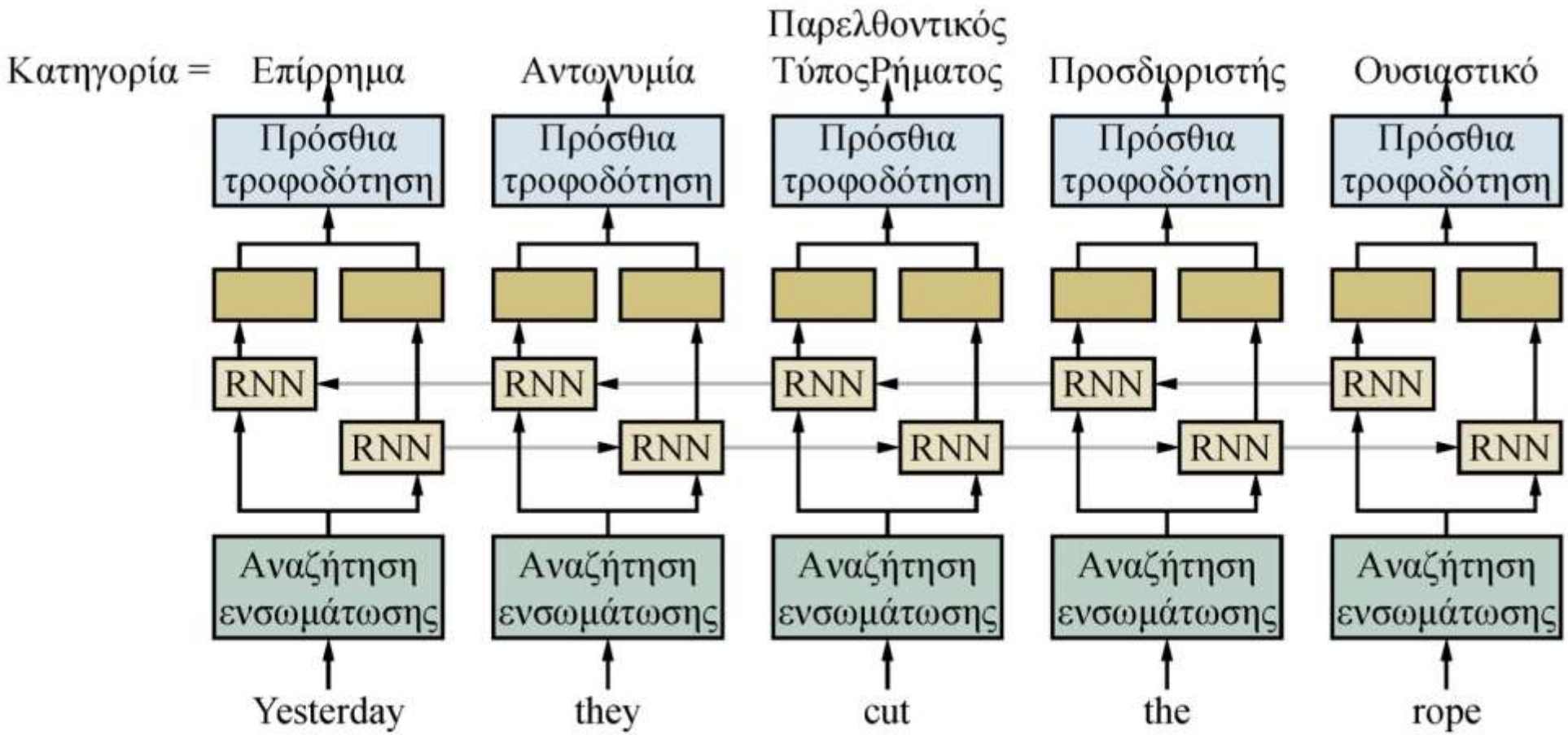
- Η αρχιτεκτονική RNN
  - λύνει το πρόβλημα των υπερβολικά πολλών παραμέτρων καθώς το πλήθος των παραμέτρων παραμένει σταθερό, ανεξάρτητα από το πλήθος των λέξεων.
  - λύνει επίσης το πρόβλημα της ασυμμετρίας, καθώς τα βάρη είναι ίδια για κάθε θέση μιας λέξης.
- Θεωρητικά δεν υπάρχει κανένα όριο ως προς το πόσο πίσω στην είσοδο μπορεί να κοιτάξει το μοντέλο. Ωστόσο, το επίπεδο z διαθέτει περιορισμένο χώρο αποθήκευσης, οπότε δεν μπορεί να «θυμάται» τα πάντα σχετικά με όλες τις προηγούμενες λέξεις.
  - Ωστόσο, η διαδικασία εκπαίδευσης ενθαρρύνει το δίκτυο να κατανείμει τον αποθηκευτικό χώρο του z σε εκείνες τις πτυχές της εισόδου που θα αποδειχθούν πραγματικά χρήσιμες.

Παραγωγή κειμένου με αναδρομικά νευρωνικά δίκτυα

- Τα γλωσσικά μοντέλα με αναδρομικά νευρωνικά δίκτυα είναι παραγωγικά μοντέλα.
- Μπορούν να παράγουν τυχαίο κείμενο μέσω μιας διαδικασίας σταθμισμένης δειγματοληψίας κάθε επόμενης λέξης, με βάση την κατανομή πιθανοτήτων επί του συνόλου των λέξεων που προκύπτει στην έξοδο του δικτύου.
- Παρακάτω φαίνεται ένα παράδειγμα τυχαίου κειμένου το οποίο έχει παραχθεί από ένα μοντέλο RNN εκπαιδευμένο στα έργα του Shakespeare (Karpathy, 2015):
  - *Marry, and will, my lord, to weep in such a one were prettiest;*
  - *Yet now I was adopted heir*
  - *Of the world's lamentable day,*
  - *To watch the next way with his father with his face?*

Ταξινόμηση με αναδρομικά νευρωνικά δίκτυα

- Τα δίκτυα RNN είναι δυνατό να χρησιμοποιούνται και για άλλες γλωσσικές εργασίες, όπως η επισημείωση με μέρη του λόγου ή η ανάλυση συναναφορών (coreference resolution).
  - Παράδειγμα συναναφοράς: «Ο Κώστας μού είπε ότι ο Ανδρέας ήταν πολύ άρρωστος, και έτσι τον πήγα στο νοσοκομείο»
- Στην περίπτωση της ταξινόμησης δεν υπάρχει κανένας λόγος να περιοριζόμαστε στην εξέταση μόνο των προηγούμενων λέξεων. Ιδιαίτερα χρήσιμο μπορεί να αποδειχθεί και το να βλέπουμε πιο μπροστά στην πρόταση.
  - Γνωρίζουμε από πειράματα οφθαλμικής ιχνηλάτησης ότι οι άνθρωποι δεν διαβάζουν πάντα από τα αριστερά στα δεξιά.
- Για να αποτυπώσουμε τα συμφραζόμενα στα δεξιά, μπορούμε να χρησιμοποιήσουμε ένα **αμφίδρομο δίκτυο RNN** (bidirectional RNN), το οποίο συνενώνει ένα ξεχωριστό μοντέλο που διαβάζει από τα δεξιά στα αριστερά με το μοντέλο που διαβάζει από τα αριστερά στα δεξιά (Στην Εικόνα 24.5).



Εικόνα 24.5 Αμφίδρομο δίκτυο RNN για επισήμανση POS.

Για ένα αμφίδρομο RNN, το *zt* θεωρείται συνήθως ως η συνένωση των διανυσμάτων από τα μοντέλα αριστερά-δεξιά και δεξιά-αριστερά.

Ταξινόμηση σε επίπεδο κειμένου

- Τα δίκτυα RNN μπορούν να χρησιμοποιούνται για εργασίες ταξινόμησης σε επίπεδο προτάσεων ή εγγράφων, όπου δεν έχουμε μια ροή εξόδων (μία έξοδο για κάθε χρονικό βήμα) αλλά μία και μοναδική έξοδο που παράγεται στο τέλος.
  - Π.χ., ανάλυση συναισθήματος (sentiment analysis)
- Η χρήση δικτύων RNN για τέτοιες εργασίες απαιτεί την παραγωγή μιας συγκεντρωτικής αναπαράστασης *y* για ολόκληρη την πρόταση από τις εξόδους μεμονωμένων λέξεων *yt* του RNN.
  - Μπορούμε να χρησιμοποιήσουμε την κρυφή κατάσταση του RNN που αντιστοιχεί στην τελευταία λέξη της εισόδου, ή
  - να συγκεντρώσουμε όλα τα κρυφά διανύσματα, π.χ., με τη **συγκέντρωση τοπικού μέσου όρου** (average pooling):

$$\tilde{\mathbf{z}} = \frac{1}{s}\sum_{t=1}^{s}\mathbf{z}_t$$

Το συγκεντρωτικό διάνυσμα διάστασης *d* μπορεί στη συνέχεια να μεταφερθεί σε ένα ή περισσότερα επίπεδα πρόσθιας τροφοδότησης, πριν τροφοδοτηθεί στο επίπεδο εξόδου.

Δίκτυα LSTM για εργασίες NLP

- Θεωρητικά, οποιαδήποτε πληροφορία θα μπορούσε να μεταβιβαστεί από ένα κρυφό επίπεδο στο επόμενο για οποιοδήποτε πλήθος χρονικών βημάτων.
- Στην πράξη, όμως, η πληροφορία μπορεί να χαθεί ή να παραμορφωθεί, όπως στο χαλασμένο τηλέφωνο.
  - –Αυτό το πρόβλημα για τα δίκτυα RNN είναι παρόμοιο με το πρόβλημα των **εξαφανιζόμενων κλίσεων** (vanishing gradients).
- Τα μοντέλα μακράς και βραχείας μνήμης (**long short-term memory, LSTM**) μπορούν να επιλέγουν να θυμούνται μερικά τμήματα της εισόδου, αντιγράφοντάς τα στο επόμενο χρονικό βήμα, και να ξεχνούν τα υπόλοιπα.
  - –Ουσιαστικά τα δίκτυα LSTM έχουν τη δυνατότητα να μαθαίνουν να δημιουργούν λανθάνοντα χαρακτηριστικά, και να αντιγράφουν αυτά τα χαρακτηριστικά προς τα εμπρός χωρίς να τα αλλοιώνουν, μέχρι να χρειαστεί να κάνουν μια τέτοια επιλογή.

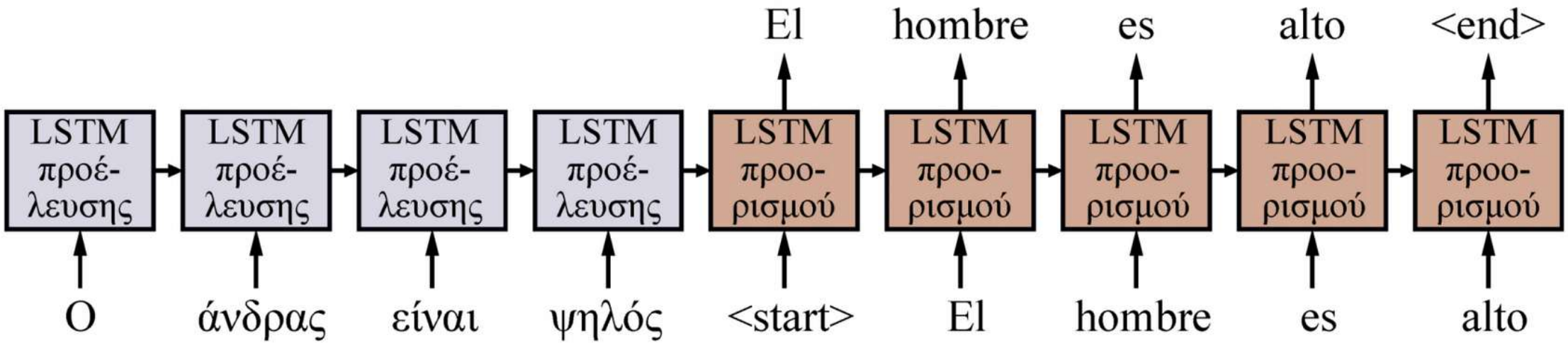
Μοντέλα ακολουθίας σε ακολουθία

Εισαγωγή

- Μία από τις ευρύτερα μελετημένες εφαρμογές της NLP είναι η **μηχανική μετάφραση** (machine translation, MT), όπου ο στόχος είναι η μετάφραση μιας πρότασης από μια **γλώσσα προέλευσης** (source language) σε μια **γλώσσα προορισμού** (target language).
- Το πρόβλημα είναι ότι δεν υπάρχει αντιστοίχιση ένα-προς-ένα μεταξύ των δύο γλωσσών.
  - Η αναδιάταξη των λέξεων μπορεί να είναι ακραία.
- Η παραγωγή κάθε λέξης προορισμού γίνεται υπό συνθήκη ως προς ολόκληρη την πρόταση προέλευσης, καθώς και όλες τις λέξεις προορισμού που έχουν παραχθεί προηγουμένως.
- Έτσι πρέπει πρώτα να διαβάσουμε ολόκληρη την πρόταση προέλευσης και μετά να αρχίσουμε να παράγουμε την πρόταση προορισμού.

Μοντέλο ακολουθίας σε ακολουθία

- Χρησιμοποιήσουμε δύο RNNs, ένα για τη γλώσσα προέλευσης και ένα για τη γλώσσα προορισμού.
- Εκτελούμε το πρώτο στην πρόταση προέλευσης, και έπειτα χρησιμοποιούμε την τελική κρυφή κατάσταση αυτού του μοντέλου ως την αρχική κρυφή κατάσταση του RNN προορισμού.
  - Με αυτόν τον τρόπο, κάθε λέξη προορισμού βρίσκεται έμμεσα υπό συνθήκη τόσο ως προς ολόκληρη την πρόταση προέλευσης όσο και προς τις προηγούμενες λέξεις προορισμού.
- Αυτή η αρχιτεκτονική νευρωνικού δικτύου ονομάζεται βασικό **μοντέλο ακολουθίας σε ακολουθία** (sequence-to-sequence model, **seq2seq**) (Εικόνα 24.6).
  - Εκτός από την μηχανική μετάφραση, μπορούν να χρησιμοποιηθούν και για άλλες εργασίες, όπως η αυτόματη δημιουργία λεζαντών κειμένου από εικόνες, ή η **σύνοψη** (summarization).



Εικόνα 24.6 Βασικό μοντέλο ακολουθίας σε ακολουθία. Κάθε ορθογώνιο αναπαριστά ένα χρονικό βήμα στο δίκτυο LSTM. (Για λόγους απλότητας, τα επίπεδα ενσωμάτωσης και εξόδου δεν εμφανίζονται.) Σε διαδοχικά βήματα τροφοδοτούμε το δίκτυο με τις λέξεις της πρότασης προέλευσης «Ο άνδρας είναι ψηλός», και έπειτα με την ετικέτα `<start>` για να υποδείξουμε ότι το δίκτυο πρέπει να αρχίσει να παράγει την πρόταση προορισμού. Η τελική κρυφή κατάσταση στο τέλος της πρότασης προέλευσης χρησιμοποιείται ως η κρυφή κατάσταση για την αρχή της πρότασης προορισμού. Μετά από αυτό, κάθε λέξη της πρότασης προορισμού σε χρόνο t χρησιμοποιείται ως είσοδος σε χρόνο $t + 1$, μέχρι το δίκτυο να παραγάγει την ετικέτα `<end>` που υποδεικνύει ότι η παραγωγή της πρότασης έχει ολοκληρωθεί.

Μειονεκτήματα μοντέλων seq2seq

- **Μεροληψία προς τα συμφραζόμενα μικρής απόστασης** (near by context bias), ειδικά σε προβλήματα με μεγάλα μεγέθη προτάσεων εισόδου, όπως η σύνοψη.
- **Σταθερό όριο μεγέθους συμφραζομένων** (fixed context size limit): Ολόκληρη η πρόταση προέλευσης συμπιέζεται σε ένα μόνο διάνυσμα κρυφής κατάστασης σταθερών διαστάσεων.
- **Βραδύτερη ακολουθιακή επεξεργασία** (slower sequential processing), λόγω της σειριακής εκπαίδευσης για κάθε πρόταση.

Προσοχή

- Στόχος είναι το RNN για τη γλώσσα προορισμού να βρίσκεται υπό συνθήκη ως προς όλα τα κρυφά διανύσματα του RNN προέλευσης, και όχι μόνο ως προς το τελευταίο.
  - –Κάτι τέτοιο θα μετρίαζε το μειονέκτημα της μεροληψίας προς τα συμφραζόμενα μικρής απόστασης, καθώς και εκείνο των σταθερών ορίων μεγέθους των συμφραζομένων, επιτρέποντας στο μοντέλο να έχει εξίσου καλή πρόσβαση σε οποιαδήποτε προηγούμενη λέξη.
- Δημιουργείται ένα «παράθυρο» κρυφών διανυσμάτων του RNN προέλευσης, από τα οποία το RNN προορισμού «μαθαίνει» να εστιάζει την **προσοχή** του (attention) σε συγκεκριμένες θέσεις κάθε φορά.
- Έστω **ci** το διάνυσμα συμφραζομένων, το οποίο περιέχει τις πιο σχετικές πληροφορίες για την παραγωγή της επόμενης λέξης στη γλώσσα προορισμού και χρησιμοποιείται ως πρόσθετη είσοδος στο RNN προορισμού.

Προσοχή

- Ένα μοντέλο ακολουθίας σε ακολουθία που χρησιμοποιεί τη συνιστώσα προσοχής ονομάζεται μοντέλο ακολουθίας σε ακολουθία με προσοχή (attentional sequence-to-sequence model).Αν ένα τυπικό RNN προορισμού γράφεται ως:

$$\mathbf{h}_i = RNN(\mathbf{h}_{i-1}, \mathbf{x}_i)$$

- ένα RNN προορισμού για μοντέλα seq2seq με προσοχή μπορεί να γραφεί ως εξής:

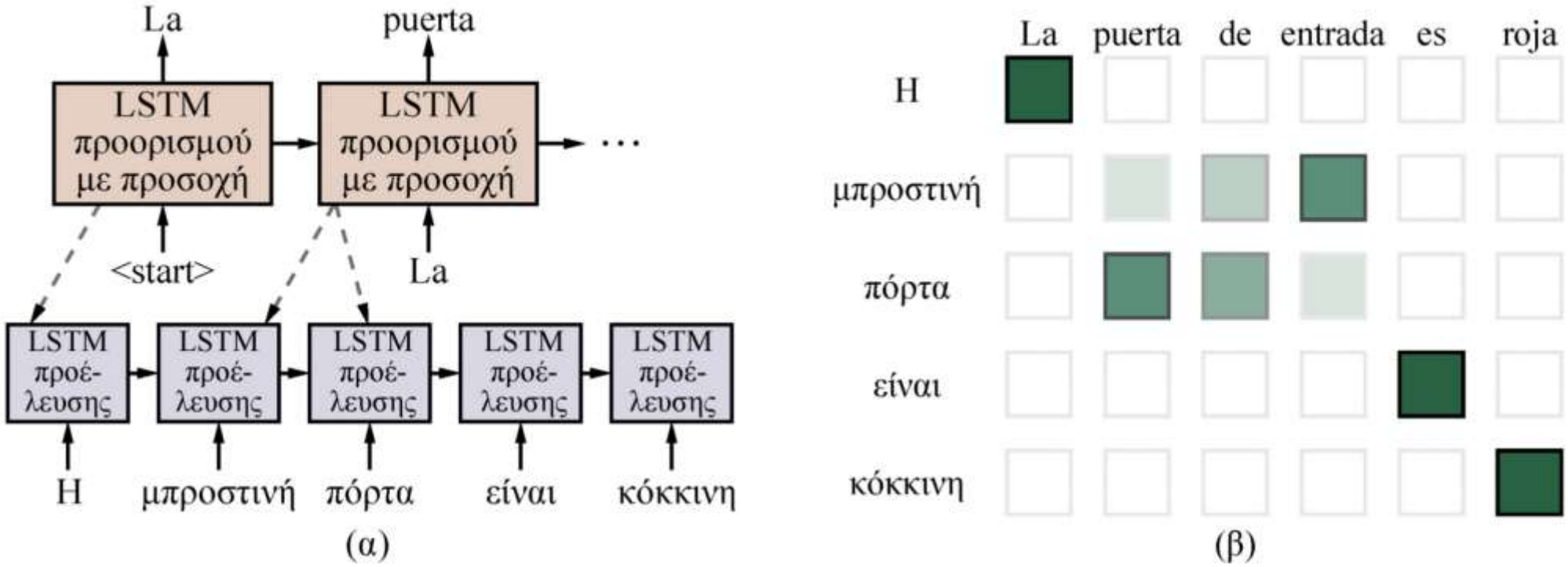
$$\mathbf{h}_i = RNN(\mathbf{h}_{i-1}, [\mathbf{x}_i; \mathbf{c}_i])$$

- όπου το [xi ; ci]είναι η συνένωση των διανυσμάτων εισόδου και συμφραζομένων, τα οποία ορίζονται ως εξής:

$$r_{ij} = \mathbf{h}_{i-1} \cdot \mathbf{s}_j \qquad a_{ij} = e^{r_{ij}} / (\sum_k e^{r_{ik}}) \qquad \mathbf{c}_i = \sum_j a_{ij} \cdot \mathbf{s}_j$$

- όπου το hi−1είναι το διάνυσμα του RNN προορισμού που πρόκειται να χρησιμοποιηθεί για την πρόβλεψη της λέξης στο χρονικό βήμα i, και το sj είναι η έξοδος του διανύσματος τουRNN προέλευσης για τη λέξη (ή το χρονικό βήμα) προέλευσης j.

- Τόσο το *hi−1* όσο και το *sj* είναι διανύσματα *d* διαστάσεων.
- Η τιμή του *rij* είναι συνεπώς η ανεπεξέργαστη «βαθμολογία προσοχής» μεταξύ της τρέχουσας κατάστασης προορισμού και της λέξης προέλευσης *j*.
  - Αυτές οι βαθμολογίες κανονικοποιούνται έπειτα σε μια πιθανότητα *aij* με χρήση μιας κατανομής *softmax* για όλες τις λέξεις προέλευσης.
- Τέλος, αυτές οι πιθανότητες χρησιμοποιούνται για την παραγωγή ενός σταθμισμένου μέσου όρου των διανυσμάτων του RNN προέλευσης, *ci* (άλλο ένα διάνυσμα *d* διαστάσεων).



Εικόνα 24.7 (α) Μοντέλο ακολουθίας σε ακολουθία με προσοχή για μετάφραση από τα Ελληνικά στα Ισπανικά. Οι διακεκομμένες γραμμές αναπαριστούν την προσοχή. (β) Παράδειγμα μήτρας πιθανοτήτων προσοχής για ένα δίγλωσσο ζεύγος προτάσεων, με τα πιο σκούρα τετράγωνα να αναπαριστούν μεγαλύτερες τιμές του a_{ij}. Οι πιθανότητες προσοχής σε κάθε στήλη έχουν άθροισμα 1.

Παρατηρήσεις

- Η συνιστώσα προσοχής δεν μαθαίνει κανένα βάρος και υποστηρίζει ακολουθίες μεταβλητού μήκους τόσο στην πλευρά της προέλευσης όσο και στην πλευρά του προορισμού.
- Η προσοχή είναι εντελώς λανθάνουσα: Ο προγραμματιστής δεν υπαγορεύει ποιες πληροφορίες θα χρησιμοποιούνται και πότε.
  - Σε πολυεπίπεδα δίκτυα RNN, η προσοχή εφαρμόζεται σε κάθε επίπεδο ξεχωριστά.
- Η πιθανοτικήδιατύπωση softmaxγια την προσοχή εξυπηρετεί τρεις σκοπούς:
  - Καθιστά την προσοχή διαφορίσιμη
  - Επιτρέπει στο μοντέλο να αποτυπώνει ορισμένους τύπους συμφραστικοποίησης μεγάλης απόστασης (long-distance contextualization)
  - Επιτρέπει στο δίκτυο να αναπαριστά την αβεβαιότητα

Παρατηρήσεις

- Οι πιθανότητες προσοχής συχνά μπορούν να ερμηνεύονται από τους ανθρώπους και να κατανοούνται διαισθητικά.
  - Στη μηχανική μετάφραση οι πιθανότητες προσοχής συχνά αντιστοιχούν στις λέξη προς λέξη ευθυγραμμίσεις που θα δημιουργούσε ένας άνθρωπος(Εικόνα 24.7(β)).

Προεκπαίδευση και μεταφορά γνώσης

Εισαγωγή

- Στο Διαδίκτυο προστίθενται καθημερινά πάνω από 100 δισεκατομμύρια λέξεις κειμένου, συμπεριλαμβανομένων ψηφιοποιημένων βιβλίων, επιμελημένων πόρων όπως η Wikipedia, και μη επιμελημένων αναρτήσεων στα μέσα κοινωνικής δικτύωσης.
- Θα προτιμούσαμε να μην χρειάζεται να δημιουργούμε νέα σύνολα δεδομένων κάθε φορά που εκπαιδεύουμε νέα μοντέλα NLP.
- Θα δούμε την ιδέα της **προεκπαίδευσης** (pretraining): πρόκειται για μια μορφή **μεταφοράς γνώσης** (transfer learning) κατά την οποία χρησιμοποιούμε μια μεγάλη ποσότητα κοινόχρηστων γλωσσικών δεδομένων γενικού περιεχομένου ώστε να εκπαιδεύσουμε μια πρώτη εκδοχή ενός μοντέλου NLP.
- Έπειτα, μπορούμε να χρησιμοποιήσουμε μικρότερες ποσότητες εξειδικευμένων δεδομένων στο πεδίο που μας ενδιαφέρει (ίσως και κάποια χαρακτηρισμένα δεδομένα) με σκοπό να βελτιστοποιήσουμε το μοντέλο σε αυτό το πεδίο.
- Το βελτιστοποιημένο μοντέλο μπορεί να μαθαίνει το λεξιλόγιο, τους ιδιωματισμούς, τις συντακτικές δομές και άλλα γλωσσικά φαινόμενα που αφορούν συγκεκριμένα το νέο πεδίο.

Το μοντέλο ενσωμάτωσης λέξεων GloVe

- GloVe (Global Vectors)
- Αρχικά συλλέγουμε μετρήσεις για το πόσες φορές εμφανίζεται κάθε λέξη μέσα σε ένα παράθυρο μιας άλλης λέξης, όπως και στο γραμμικό μοντέλο παράλειψης (skip-gram).
- Πρώτα επιλέγουμε το μέγεθος παραθύρου (πιθανόν 5 λέξεις) και ορίζουμε ότι *Xij* είναι το πλήθος των φορών που οι λέξεις *i* και *j* συνεμφανίζονται στο παράθυρο, και *Xi* είναι το πλήθος των φορών που η λέξη *i* συνεμφανίζεται με οποιαδήποτε άλλη λέξη.
- Η σημειογραφία *Pij = Xij/Xi* υποδεικνύει την πιθανότητα να εμφανίζεται η λέξη *I* στα συμφραζόμενα της λέξης *j*.
- Όπως και πριν, το *Ei* συμβολίζει την ενσωμάτωση της λέξης *i*

Το μοντέλο GloVe

- Μέρος της φιλοσοφίας του μοντέλου GloVe είναι ότι η σχέση μεταξύ δύο λέξεων μπορεί να αποτυπωθεί καλύτερα μέσω της σύγκρισης και των δύο με άλλες λέξεις.
- Έστω, για παράδειγμα, οι λέξεις πάγος και ατμός. Ας εξετάσουμε, λοιπόν, τον λόγο των πιθανοτήτων τους να συνεμφανίζονται με κάποια άλλη λέξη w, που είναι:

$$P_{w,\text{πάγος}} / P_{w,\text{ατμός}}$$

  - Όταν το w είναι η λέξη στερεό, ο λόγος θα είναι μεγάλος
  - Όταν το w είναι η λέξη αέριο, ο λόγος θα είναι μικρός
  - Όταν το w είναι μια λέξη εξίσου άσχετη ή εξίσου σχετική και με τις δύο, ο λόγος θα είναι κοντά στο 1.
- Στο μοντέλο GloVe, το εσωτερικό γινόμενο δύο διανυσμάτων λέξεων είναι ίσο με τη λογαριθμική πιθανότητα της συνεμφάνισής τους.

Μια τεχνική περιπλοκότητα είναι ότι το μοντέλο GloVe δημιουργεί δύο διανύσματα ενσωμάτωσης για κάθε λέξη. Ο υπολογισμός των δύο και η άθροισή τους στο τέλος βοηθά στον περιορισμό της υπερπροσαρμογής

Σύγχρονη τεχνολογία

Η τρέχουσα κατάσταση

- Η μεταφορά γνώσης αποδίδει καλά στα προβλήματα φυσικής γλώσσας: μπορούμε να πάρουμε ένα γενικό γλωσσικό μοντέλο από τον Ιστό και να το βελτιστοποιήσουμε για τις ανάγκες μιας συγκεκριμένης εργασίας.
  - Όλα ξεκίνησαν με απλές ενσωματώσεις λέξεων από συστήματα όπως το WORD2VEC (2013) και το GloVe (2014).
- Τα μοντέλα αυτά έγιναν εφικτά μόνο αφότου έγινε ευρεία η χρήση προηγμένων τεχνολογιών υλικού (GPU και TPU), με τους ερευνητές να είναι ευγνώμονες που μπορούσαν να κατεβάζουν μοντέλα αντί να χρειάζεται να δαπανούν πόρους για να εκπαιδεύουν δικά τους.
- Το μοντέλο μετασχηματιστή επέτρεψε την αποδοτική εκπαίδευση πολύ μεγαλύτερων και βαθύτερων νευρωνικών δικτύων από ό,τι ήταν προηγουμένως δυνατό (αυτή τη φορά με την πρόοδο του λογισμικού, όχι του υλικού).
- Από το 2018 τα νέα έργα NLP συνήθως ξεκινούν με ένα προεκπαιδευμένο μοντέλο μετασχηματιστή.

Το σύστημα GPT-2

- Αν και αυτά τα μοντέλα εκπαιδεύονται να προβλέπουν την επόμενη λέξη στο κείμενο, αποδίδουν εντυπωσιακά καλά και σε άλλες γλωσσικές εργασίες.
- Το GPT2, ένα γλωσσικό μοντέλο τύπου μετασχηματιστή με 1,5 δισεκατομμύριο παραμέτρους εκπαιδευμένες σε 40 GB διαδικτυακού κειμένου, επιτυγχάνει καλά αποτελέσματα σε πολύ διαφορετικές εργασίες, όπως η μετάφραση Γαλλικών-Αγγλικών, η εύρεση των λέξεων αναφοράς σε εξαρτήσεις μεγάλης απόστασης, και η απάντηση σε ερωτήσεις γενικών γνώσεων, χωρίς να έχει προσαρμοστεί επακριβώς στις συγκεκριμένες εργασίες.
  - Όπως φαίνεται στην Εικόνα 24.14, το GPT2 μπορεί να παράγει αρκετά φυσικό κείμενο με δεδομένες μόνο λίγες λέξεις ως έναυσμα.

It is a truth universally acknowledged that the earth is not the center of the universe. There are those who assert there is. I do not accept them, but others I consider to be of the same opinion. The truth is, however, that if there are other than the center, and if there are any other living things in the universe and if they are not human, then we do not yet have our answers. We have to go on. This page gives a simplified, simplified answer to the problem. We don't have all the answers. The truth is, however, that the truth is out there.

When Gregor Samsa woke up one morning, he did not notice anything strange. "When my wife is looking at me, I feel like she is looking at a piece of art," he said. "I think she is admiring something I have created." The idea is that by looking at your own life, you learn something important and become a better person. It is a theory that emerged from psychologist Daniel Goleman's work, in which he asked "How do you know you're not a loser?"

Εικόνα 24.14 Παραδείγματα συμπλήρωσης κειμένου που έχουν παραχθεί από το γλωσσικό μοντέλο GPT-2, με δεδομένα τα εναύσματα σε **έντονη γραφή**. Τα περισσότερα κείμενα είναι σε σχεδόν άπταιστα Αγγλικά, τουλάχιστον σε τοπικό επίπεδο. Στο τελευταίο παράδειγμα φαίνεται ότι κάποιες φορές το μοντέλο απλώς καταρρέει.

Alice was beginning to get very tired of sitting with her sister on the bank. She sat up, yawned, and said, with a loud little scream, "I hope you don't mind if I keep on doing what I should like to do, and if someone asks me which of us will do more, don't tell them that I won't do much, my dear sister."

All happy families are alike; each happy family is like a garden of paradise. The only difference between happy families and unhappy families, is that the unhappy family doesn't have any flowers or trees.

Tell me a story. Tell me a story. Tell me a story. Tell me a story. Tell me a story. Tell me a story. Tell me a story. Tell me a story. Tell me a story. Tell me a story. Tell me a story. Please fill out the following details. Thank you... Thank you for your interest in this interview. Please wait...

Εικόνα 24.14 Παραδείγματα συμπλήρωσης κειμένου που έχουν παραχθεί από το γλωσσικό μοντέλο GPT-2, με δεδομένα τα εναύσματα σε **έντονη γραφή**. Τα περισσότερα κείμενα είναι σε σχεδόν άπταιστα Αγγλικά, τουλάχιστον σε τοπικό επίπεδο. Στο τελευταίο παράδειγμα φαίνεται ότι κάποιες φορές το μοντέλο απλώς καταρρέει.

Το σύστημα ARISTO (Clark κ.ά., 2019)

- Το σύστημα ARISTO, ένα χαρακτηριστικό παράδειγμα προηγμένου συστήματος NLP, πέτυχε βαθμολογία 91,6% σε ένα διαγώνισμα πολλαπλής επιλογής στη φυσική της δευτέρας γυμνασίου (Εικόνα 24.13).
- Το ARISTOαποτελείται από μια συλλογή επιλυτών: κάποιοι από αυτούς χρησιμοποιούν ανάκτηση πληροφοριών, κάποιοι εφαρμόζουν κειμενική συνεπαγωγή και ποιοτική συλλογιστική, και κάποιοι χρησιμοποιούν μεγάλα γλωσσικά μοντέλα μετασχηματιστή.
- Το ARISTO πέτυχε βαθμολογία 83% στο πιο προχωρημένο διαγώνισμα της τρίτης λυκείου.
- Ωστόσο, το ARISTOπάσχει από ορισμένους περιορισμούς. Ασχολείται μόνο με ερωτήσεις πολλαπλής επιλογής, όχι με ελεύθερης ανάπτυξης, και δεν μπορεί ούτε να διαβάζει ούτε να παράγει διαγράμματα.

1. **What will best separate a mixture of iron filings and black pepper?**
   (a) magnet (b) filter paper (c) triple beam balance (d) voltmeter
2. **Which form of energy is produced when a rubber band vibrates?**
   (a) chemical (b) light (c) electrical (d) sound
3. **Because copper is a metal, it is**
   (a) liquid at room temperature (b) nonreactive with other substances
   (c) a poor conductor of electricity (d) a good conductor of heat
4. **Which process in an apple tree primarily results from cell division?**
   (a) growth (b) photosynthesis (c) gas exchange (d) waste removal

Εικόνα 24.13 Ερωτήσεις από διαγώνισμα στη φυσική, στις οποίες το σύστημα ARISTO μπορεί να απαντά σωστά χρησιμοποιώντας μια συλλογή από μεθόδους, με πιο σημαντική το γλωσσικό μοντέλο ROBERTA. Η απάντηση αυτών των ερωτήσεων προϋποθέτει τη γνώση της φυσικής γλώσσας, της δομής των διαγωνισμάτων πολλαπλής επιλογής, της κοινής λογικής, και της επιστήμης.

Μελλοντικές κατευθύνσεις

- Υπολείπεται ακόμα πολλή δουλειά για τη βελτίωση των συστημάτων NLP.
- Ένα πρόβλημα είναι ότι τα μοντέλα μετασχηματιστή βασίζονται μόνο στα κοντινά συμφραζόμενα, τα οποία περιορίζονται σε μερικές εκατοντάδες λέξεις.
- Τα μοντέλα που βασίζονται σε δεδομένα αναπτύσσονται και συντηρούνται ευκολότερα, και επιτυγχάνουν καλύτερες βαθμολογίες σε καθιερωμένες μετρήσεις, σε σύγκριση με τα συστήματα που κατασκευάζονται «με το χέρι» και τα οποία προϋποθέτουν την εφαρμογή των προσεγγίσεων που αναφέραμε καθώς επίσης και αρκετή ανθρώπινη προσπάθεια.
- Δεν αποκλείεται ωστόσο οι μελλοντικές ανακαλύψεις στη ρητή γραμματική και σημασιολογική μοντελοποίηση να γείρουν τη ζυγαριά υπέρ των μοντέλων που κατασκευάζονται «με το χέρι».
- Ίσως είναι πιθανότερη η εμφάνιση υβριδικών προσεγγίσεων που συνδυάζουν τις καλύτερες έννοιες και από τους δύο κόσμους.