ΓΛΩΣΣΙΚΗ ΤΕΧΝΟΛΟΓΙΑ

ΕΞΑΓΩΓΗ ΠΛΗΡΟΦΟΡΙΑΣ

INFORMATION EXTRACTION

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

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

Extraction)

- Δημιουργία μιας δομημένης αναπαράστασης (π.χ. μία βάση δεδομένων) επιλεγμένης πληροφορίας που εξάγεται από κείμενα φυσικής γλώσσας
- Από τις πιο ενεργές περιοχές εφαρμοσμένης έρευνας στη γλωσσική τεχνολογία

Τρομοκρατικές Ενέργειες

- Δίνεται μία συλλογή κειμένων γύρω από ειδήσεις τρομοκρατικών ενεργειών
- Για κάθε κείμενο καθόρισε:
  - Τον τύπο της ενέργειας
  - Την ημερομηνία
  - Την τοποθεσία
  - κτλ.
- Συμπλήρωσε μία βάση δεδομένων με αυτά τα στοιχεία (templates)

Τρομοκρατικές Ενέργειες

Είσοδος:

- 19 March - A bomb went off this morning near a power tower in San Salvador leaving a large part of the city without energy, but no casualties have been reported. According to unofficial sources, the bomb - allegedly detonated by urban guerrilla commandos - blew up a power tower in the northwestern part of San Salvador at 0650 (1250 GMT).

Τρομοκρατικές Ενέργειες

Έξοδος:

- Incident type	bombing
- Date	March 19
- Location	El Salvador: San Salvador (city)
- Perpetrator	urban guerilla commandos
- Physical target	power tower
- Human target	-
- Effect on physical target	destroyed
- Effect on human target	no injury or death
- Instrument	bomb

(MUCs)

- Η έρευνα στο χώρο της εξαγωγής πληροφορίας στηρίχτηκε πολύ από τα MUCs
  - Καθορισμός πεδίου (domain) κειμένων
  - Καθορισμός κειμένων εκπαίδευσης
  - Καθορισμός προτύπων εξαγόμενων δεδομένων (templates)
  - Καθορισμός μεθόδου αξιολόγησης συστημάτων
- Τα συμμετέχοντα συστήματα ανταγωνίζονται μεταξύ τους

Σύντομη Ανασκόπηση MUCs

- MUC-1 (1987):
  - Domain: tactical naval operations reports
  - Texts: 12 for training, 2 for testing
  - 6 systems participated
- MUC-2 (1989):
  - Domain: the same
  - Texts: 105 messages for training, 25 for training)
  - 8 systems participated
- MUC-3 (1991);
  - Domain: newswire stories about terrorist attacks in nine Latin American countries
  - 1300 development texts were supplied; three test sets of 100 texts each
  - 15 systems participated
- MUC-4 (1992);
  - Domain: the same
  - different task definition and corpus etc.
  - 17 systems participated

Σύντομη Ανασκόπηση MUCs

- MUC-5 (1993)
  - 2 domains: joint ventures in financial newswire stories and microelectronics products announcements
  - 2 languages (English and Japanese)
  - 17 systems participated
  - Larger corpora
- MUC-6 (1995);
  - Domain: management succession events in financial news stories
  - Several subtasks
  - 17 systems participated
- MUC-7 (1998);
  - Domain: air vehicle (airplane, satellite,...) launch reports
  - (http://www.itl.nist.gov/iaui/894.02/related_projects/muc/)

Ανάκτηση και Εξαγωγή Πληροφορίας

- Δεδομένης μιας συλλογής κειμένων:
  - Η **ανάκτηση** πληροφορίας **(information retrieval)** επιλέγει ένα υποσύνολο κειμένων που (πιθανότατα) έχουν σχέση με ένα θέμα ή μία ερώτηση του χρήστη
  - Η **εξαγωγή** πληροφορίας **(information extraction)** θεωρεί ότι όλα τα κείμενα είναι σχετικά με το θέμα και εξάγει σχετική πληροφορία από τα κείμενα
- Η ανάκτηση και η εξαγωγή πληροφορίας είναι συμπληρωματικές τεχνολογίες

Κατανόηση Κειμένου

- Στην εξαγωγή πληροφορίας
  - Δεν απαιτείται πλήρης κατανόηση του κειμένου
  - Δεν απαιτείται πλήρης συντακτική-σημασιολογική ανάλυση
  - Υπάρχει περιορισμός θέματος, ύφους
  - Τα συστήματα είναι αξιόπιστα και μπορούν να χειριστούν ακόμα και λάθος προτάσεις
  - Τα συστήματα μπορούν να αξιολογηθούν εύκολα

Κατανόηση Κειμένου

- Στην κατανόηση κειμένου
  - Ο στόχος είναι να κατανοηθεί ακόμα και η παραμικρή λεπτομέρεια
  - Απαιτείται πλήρης συντακτική-σημασιολογική ανάλυση
  - Όλη η επεξεργασία πρέπει να είναι ανεξάρτητη πεδίου (domain-independent)
  - Πρέπει να αναγνωριστούν οι στόχοι του συγγραφέα, το ύφος, κτλ.

Ορολογία

- Domain (πεδίο)
  - Γενική θεματική περιοχή (π.χ. οικονομικά νέα)
- Scenario (σενάριο)
  - Καθορισμός των συγκεκριμένων γεγονότων ή των σχέσεων που θα εξαχθούν (π.χ. κοινοπραξίες)
- Template (φόρμα)
  - Τελική μορφή εξόδου προς συμπλήρωση από ένα σύστημα εξαγωγής πληροφορίας
- Template slot (όρισμα φόρμας)
  - Όρισμα ενός template που δέχεται μία τιμή

Γενική Αρχιτεκτονική

- Τα πιο πολλά συστήματα ακολουθούν την εξής διαδικασία:
  - Εξαγωγή ανεξάρτητων οντοτήτων/γεγονότων από το κείμενο μέσω τοπικής ανάλυσης
  - Ολοκλήρωση αυτών των οντοτήτων/γεγονότων και παραγωγή μεγαλύτερων νέων οντοτήτων/γεγονότων (μέσω συμπερασμού)
  - Τα γεγονότα ολοκληρώνονται και μεταφράζονται στη μορφή της εξόδου που απαιτείται από τις προδιαγραφές

Φάσεις Ανάλυσης

- Λεξιλογική ανάλυση
- Αναγνώριση ονομάτων-οντοτήτων
- Συντακτική ανάλυση
- Ταίριασμα προτύπου σεναρίου
- Ανάλυση συν-αναφοράς
- Συμπερασμός και συγχώνευση γεγονότων

Παράδειγμα: Διάσχιση Στελεχών Επιχειρήσεων (MUC-6)

"Sam Schwartz retired as executive vice president of the famous hot dog manufacturer, Hupplewhite Inc. He will be succeeded by Harry Himmelfarb."

Frame (Template) Example

"Sam Schwartz retired as executive vice president of the famous hot dog manufacturer, Hupplewhite Inc. He will be succeeded by Harry Himmelfarb."

- **Event** leave job
- **Person** Sam Schwartz
- **Position** executive vice president
- **Company** Hupplewhite Inc.

- **Event** start job
- **Person** Harry Himmelfarb
- **Position** executive vice president
- **Company** Hupplewhite Inc.

- Το κείμενο χωρίζεται σε λέξεις (tokens) και προτάσεις
- Κάθε λέξη χαρακτηρίζεται μορφο-συντακτικά
  - POS Tagger
  - Αναζήτηση σε λεξικό
- Λεξικά
  - Γενικά
  - Ειδικά (κυριότερες ονομασίες τόπων, εταιρειών, κυρίων ονομάτων, συντομεύσεων)

(Named Entity Recognition)

- Αναγνωρίζονται διάφοροι τύποι κύριων ονομάτων και άλλες ειδικές μορφές οντοτήτων
  - (π.χ. ημερομηνίες, ποσά, διευθύνσεις)
- Οι ονοματικές οντότητες (named-entities) εμφανίζονται συχνά σε πολλά είδη κειμένων
- Η αναγνώριση και η ταξινόμησή τους διευκολύνει την περαιτέρω επεξεργασία
- Οι ονοματικές οντότητες είναι σημαντικές γιατί συχνά καλύπτουν τιμές των template slots

Δυσκολίες

- Κανένα λεξικό δεν μπορεί να συμπεριλάβει όλα τα υπάρχοντα κύρια ονόματα, διευθύνσεις κτλ.
- Νέα κύρια ονόματα δημιουργούνται συνεχώς
- Η ίδια ονοματική οντότητα μπορεί να αναφέρεται με διάφορες παραλλαγές (Coca Cola - Coke)
- Συνήθως δημιουργούνται ακρωνύμια- συντομεύσεις για τα κύρια ονόματα
- Τα ακρωνύμια-συντομεύσεις δεν είναι πάντα κύρια ονόματα (π.χ. λ.χ. κτλ.)
- Η χρήση κεφαλαίων γραμμάτων δεν είναι πάντα κανόνας

Ασάφεια Ονομάτων - Οντοτήτων

- Ακόμα και αν αναγνωριστεί ένα όνομα-οντότητα συχνά είναι δύσκολο να ταξινομηθεί σωστά
  - Άνθρωπος ή εταιρεία: *Ford*, *Philip Morris*
  - Άνθρωπος ή τοποθεσία: *Jordan*, *JFK*
  - Άνθρωπος ή μήνας: *April*, *June*
  - Ακρωνύμιο ή οργανισμός: MRI (Magnetic Resonance Imaging, Mental Research Institute)
  - ...

Προσεγγίσεις

- Χειρονακτικοί κανόνες
  - Συνήθως αποδίδουν καλύτερα για εξειδικευμένες εφαρμογές
  - Δύσκολο να κατασκευαστούν
  - Domain-specific (Εξαρτώμενοι από την θεματική περιοχή)
- Μηχανική μάθηση
  - Μπορεί να προσαρμοστεί εύκολα σε νέα πεδία
  - Χρειάζεται domain-specific δεδομένα εκπαίδευσης
- Η αναγνώριση ονοματικών οντοτήτων είναι από τις πιο πετυχημένες εργασίες επεξεργασίας φυσικής γλώσσας (ακρίβεια ~95%)

Χειρονακτικοί Κανόνες

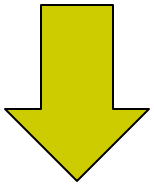
- Βασίζονται σε
  - Ειδικά λεξικά (λίστες ονομάτων, εταιρειών, κτλ)
  - Λέξεις κλειδιά (Mr., Jr., Ph.D., Corp., Inc., Co.,)
  - Πρότυπα συμφραζομένων
    - `{TITLE} {PERSON}`
      `Mr. Frank Leonard`
    - `{TITLE}, the {TITLE} of {ORGANIZATION}`
      `Fred Martin, the CEO of XYZ Corp.`
    - `{PERSON} joined {COMPANY}`
      `Mary Smith joined Microsoft`
    - `{LOCATION} , {LOCATION}`
      `Salt Lake City, Utah`

Εύρεση Συνωνύμων

- Υποπρόβλημα: εύρεση όλων των δυνατών συνώνυμων για μία ονοματική οντότητα
  - Larry Liggett = Mr. Liggett
  - Hewlett-Packard Corp. = HP
- Μπορεί να βοηθήσει την ταξινόμηση των οντοτήτων σε κατηγορίες
  - “Humble Hopp reported…” (person or company?)
  - “Mr. Hopp” (-> person)

Αναγνώριση Ονομάτων-Οντοτήτων

"Sam Schwartz retired as executive vice president of the famous hot dog manufacturer, Hupplewhite Inc. He will be succeeded by Harry Himmelfarb."



<name type="person"> **Sam Schwartz** </name>
retired as executive vice president of the famous hot dog manufacturer,
<name type="company"> **Hupplewhite Inc.** </name>
He will be succeeded by
<name type="person"> **Harry Himmelfarb** </name>.

Συντακτική Ανάλυση

- Η αναγνώριση της πλήρους συντακτικής ανάλυσης μιας πρότασης είναι δύσκολη
- Η αναγνώριση μερικών συστατικών της συντακτικής δομής συχνά αρκεί στα πλαίσια της εξαγωγής πληροφορίας
  - Οι ονοματικές φράσεις συχνά αντιστοιχούν στα ορίσματα των templates
  - Η σχέση κύριο ρήμα – υποκείμενο – αντικείμενο συχνά αρκεί για την κατανόηση της ενέργειας

Συστήματα Εξαγωγής Πληροφορίας

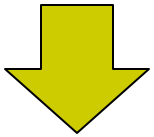
- Το κάθε σύστημα δίνει στη συντακτική επεξεργασία τη δική του βαρύτητα
  - Μερικά συστήματα δεν έχουν ξεχωριστή φάση συντακτικής επεξεργασίας
  - Άλλα συστήματα προσπαθούν να παράγουν ένα πλήρες συντακτικό δέντρο ανάλυσης
  - Τα περισσότερα συστήματα κάνουν κάτι ενδιάμεσο και παράγουν κομμάτια του δέντρου ανάλυσης (μερική συντακτική ανάλυση - partial parsing)

(Partial Parsing)

- Τα συστήματα που βασίζονται σε partial parsing κατασκευάζουν δομές για τις οποίες μπορούν να είναι σίγουρα είτε από συντακτικής είτε από σημασιολογικής άποψης
  - Για παράδειγμα, απλές ονοματικές φράσεις (άρθρο επίθετο ουσιαστικό) και ρηματικές φράσεις (ρήμα και βοηθητικά ρήματα) μπορούν να ανιχνευτούν εύκολα
  - Μεγαλύτερες δομές (πιο σύνθετες ονοματικές – ρηματικές φράσεις) μπορούν να κατασκευαστούν αν υπάρχει αρκετή σημασιολογική πληροφορία

Αναγνώριση Ονοματικών Ρηματικών Φράσεων

<name type="person"> Sam Schwartz </name>
retired as executive vice president of the famous hot dog manufacturer,
<name type="company"> Hupplewhite Inc. </name>
He will be succeeded by
<name type="person"> Harry Himmelfarb </name>.



<np entity="e1"> Sam Schwartz </np>
<vg> retired </vg>
as
<np entity="e2"> executive vice president </np>
of
<np entity="e3"> the famous hot dog manufacturer </np>,
<np entity="e4"> Hupplewhite Inc. </np>
<np entity="e5"> He </np>
<vg> will be succeeded </vg>
by
<np entity="e6"> Harry Himmelfarb </np>.

Μορφο-συντακτικές ιδιότητες

- Σε κάθε οντότητα που αναγνωρίζεται αποδίδεται μορφο-συντακτική πληροφορία ανάλογα με το είδος της
  - Στα noun groups: αριθμός, όνομα, λήμμα
  - Στα verb groups: χρόνος, φωνή, λήμμα
- Αυτή η πληροφορία μπορεί να χρησιμοποιηθεί στα επόμενα στάδια

Σημασιολογικές Ιδιότητες

Σε κάθε ονοματική φράση αποδίδεται ένας σημασιολογικός τύπος

```
<np entity="e1"> Sam Schwartz </np>
<vg> retired </vg>
as
<np entity="e2"> executive vice president </np>
of
<np entity="e3"> the famous hot dog
manufacturer </np>,
<np entity="e4"> Hupplewhite Inc. </np>
<np entity="e5"> He </np>
<vg> will be succeeded </vg>
by
<np entity="e6"> Harry Himmelfarb </np>.
```

- e1 type: person name: "Sam Schwartz"
- e2 type: position value: "executive vice president"
- e3 type: manufacturer
- e4 type: company name:"Hupplewhite Inc."
- e5 type: person
- e6 type: person name: "Harry Himmelfarb"

Αναγνώριση πιο Σύνθετων Δομών

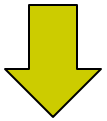
- Για να κατασκευαστούν πιο σύνθετες δομές πρέπει στις ήδη υπάρχουσες φράσεις να προσαρτηθούν συστατικά από τα δεξιά
- Υπάρχει μεγάλη ασάφεια
- Χρησιμοποιούνται σημασιολογικοί περιορισμοί για να μειώσουν την ασάφεια

Πρότυπα Αναγνώρισης

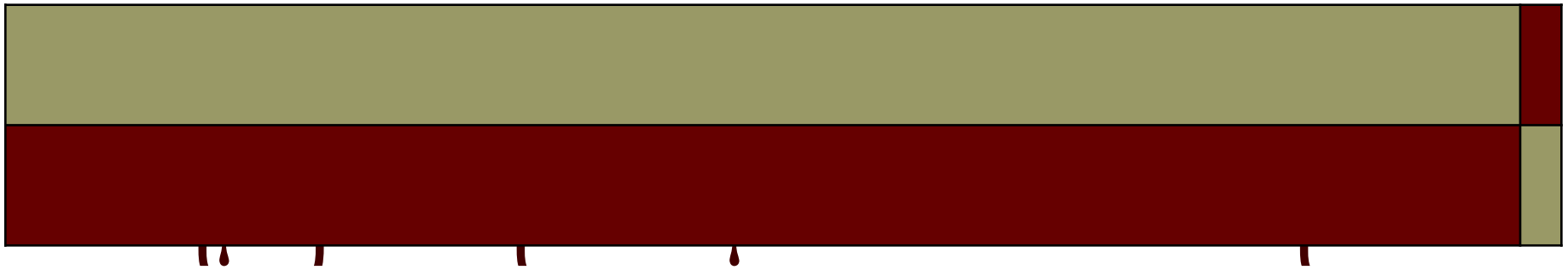
- {company-description}
  - NP, type: company, head: common noun
- {company-name}
  - NP, type: company, head: name
- {position} of {company}
  - NP, type: position
  - NP, type: company
- Μία ονοματική φράση από αυτές που έχουμε βρει ως τώρα μπορεί να ταιριάξει με μία θέση στο πρότυπο αν ταιριάζει ο σημασιολογικός της τύπος
- Υπάρχει μία μικρή *is-a* ιεραρχία σημασιολογικών τύπων
  - manufacturer → company

Anaphoric Coreference Example

<np entity="e1"> Sam Schwartz </np>
<vg> retired </vg>
as
<np entity="e2"> executive vice president </np>
of
<np entity="e3"> the famous hot dog manufacturer </np>,
<np entity="e4"> Hupplewhite Inc. </np>
<np entity="e5"> He </np>
<vg> will be succeeded </vg>
by
<np entity="e6"> Harry Himmelfarb </np>.



<np entity="e1"> Sam Schwartz </np>
<vg> retired </vg>
as
<np entity="e2"> executive vice president of the famous hot dog manufacturer, Hupplewhite Inc. </np>
<np entity="e5"> He </np>
<vg> will be succeeded </vg>
by
<np entity="e6"> Harry Himmelfarb </np>.



- e1 type: person name: "Sam Schwartz"
- e2 type: position value: "executive vice president"
- e3 type: manufacturer
- e4 type: company name:"Hupplewhite Inc."
- e5 type: person
- e6 type: person name: "Harry Himmelfarb"

- entity e1 type: person name: "Sam Schwartz"
- entity e2 type: position value: "executive vice president"
  company: e3
- entity e3 type: manufacturer name: "Hupplewhite Inc."
- entity e5 type: person
- entity e6 type: person name: "Harry Himmelfarb"

- Ο ρόλος των προτύπων σεναρίων είναι να εξάγουν τα γεγονότα και τις σχέσεις που είναι σχετικές με το σενάριο
- Στο παράδειγμά μας, υπάρχουν δύο πρότυπα
  - {person} retires as {position}
  - {person} is succeeded by {person}
- {person} και {position} είναι στοιχεία του προτύπου που ταιριάζουν με NPs με αντίστοιχο σημασιολογικό τύπο
- “retires” και “is succeeded” είναι στοιχεία του προτύπου που ταιριάζουν με ρηματικές φράσεις ενεργητικής και παθητικής φωνής, αντίστοιχα

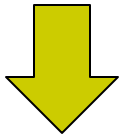
Ταίριασμα Προτύπου Σεναρίου

- Το κείμενο αποτελείται από 2 προτάσεις (clauses)
  - Η κάθε πρόταση δείχνει μία δομή γεγονότων
  - Οι δομές γεγονότων δείχνουν τις οντότητες

```
<clause event="e7"> Sam Schwartz
retired as executive vice president
of the famous hot dog manufacturer,
Hupplewhite Inc. </clause>
<clause event="e8"> He will be
succeeded by  Harry Himmelfarb
</clause>.
```

Γεγονότων

- entity e1 type: person name: "Sam Schwartz"
- entity e2 type: position value: "executive vice president" company: e3
- entity e3 type: manufacturer name: "Hupplewhite Inc."
- entity e5 type: person
- entity e6 type: person name: "Harry Himmelfarb"



- entity e1 type: person name: "Sam Schwartz"
- entity e2 type: position value: "executive vice president" company: e3
- entity e3 type: manufacturer name: "Hupplewhite Inc."
- entity e5 type: person
- entity e6 type: person name: "Harry Himmelfarb"

- event e7 type: leave-job person: e1 position: e2
- event e8 type: succeed person1: e6 person2: e5

Coreference resolution)

- Αποσαφήνιση παραπομπών από αντωνυμίες και οριστικές ονοματικές φράσεις
  - Στο παράδειγμά μας: “he” (οντότητα e5)
- Λίστα ιστορίας
  - Ψάχνουμε την πιο πρόσφατα αναφερθείσα οντότητα τύπου *person*, και βρίσκουμε την e1
  - Αναφορές στην e5 αλλάζουν και αναφέρονται στην e1
- Χρησιμοποιείται επίσης η ιεραρχία *is-a*

Γεγονότων

- entity e1 type: person name: "Sam Schwartz"
- entity e2 type: position value: "executive vice president"
- company: e3
- entity e3 type: manufacturer name: "Hupplewhite Inc."
- entity e5 type: person
- entity e6 type: person name: "Harry Himmelfarb"

- event e7 type: leave-job person: e1 position: e2
- event e8 type: succeed person1: e6 person2: e5

- entity e1 type: person name: "Sam Schwartz"
- entity e2 type: position value value: "executive vice president"
- company: e3
- entity e3 type: manufacturer name:"Hupplewhite Inc."
- entity e6 type: person name: "Harry Himmelfarb"

- event e7 type: leave-job person: e1 position: e2
- event e8 type: succeed person1: e6 person2: e1

Γεγονότων

- Ένα γεγονός μπορεί να περιγράφεται σε πολλές προτάσεις
  - Αυτή η πληροφορία πρέπει να συνδυαστεί πριν δημιουργηθεί ένα template
- Κάποια πληροφορία μπορεί να μην αναφέρεται ρητώς
  - Αυτή η πληροφορία πρέπει να γίνει σαφής μέσω μιας διαδικασίας συμπερασμού

Γεγονότων

- Στο παράδειγμά μας, πρέπει να καθορίσουμε τι υπονοεί το ρήμα “succeed”
  - Παράδειγμα 1: “Sam was president. He was succeeded by Harry.”

    Συμπέρασμα: Harry will become president

  - Παράδειγμα 2: “Sam will be president. He succeeds Harry”

    Συμπέρασμα: Harry was president.

Γεγονότων

- Τέτοια συμπεράσματα μπορούν να προκύψουν από κανόνες συστημάτων παραγωγής:

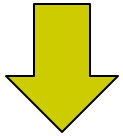
```
leave-job(X-person,Y-job) &
succeed(Z-person,X-person)
=> start-job(Z-person,Y-job)

start-job(X-person,Y-job) &
succeed(X-person,Z-person)
=>leave-job(Z-person,Y-job)
```

Γεγονότων

- entity e1 type: person name: "Sam Schwartz"
- entity e2 type: position value: "executive vice president"
- company: e3
- entity e3 type: manufacturer name:"Hupplewhite Inc."
- entity e6 type: person name: "Harry Himmelfarb"

- event e7 type: leave-job person: e1 position: e2
- event e8 type: succeed person1: e6 person2: e1



- entity e1 type: person name: "Sam Schwartz"
- entity e2 type: position value: "executive vice president"
- company: e3
- entity e3 type: manufacturer name:"Hupplewhite Inc."
- entity e6 type: person name: "Harry Himmelfarb"

- event e7 type: leave-job person: e1 position: e2
- event e8 type: succeed person1: e6 person2: e1
- event e9 type: start-job person: e6 position:e2

Poppa's Example (Target Templates)

- **Event** leave job
- **Person** Sam Schwartz
- **Position** executive vice president
- **Company** Hupplewhite Inc.

- **Event** start job
- **Person** Harry Himmelfarb
- **Position** executive vice president
- **Company** Hupplewhite Inc

- Το απλό σενάριο του παραδείγματος δεν απαιτούσε να λάβουμε υπόψη τον χρόνο που συνέβηκε το κάθε γεγονός
- Για πολλά σενάρια, ο χρόνος είναι πολύ σημαντικός
  - Πρέπει να αναφέρονται ρητά κάποιες χρονικές τιμές
  - Η ακολουθία των γεγονότων παίζει σημαντικό ρόλο
- Η πληροφορία για τον χρόνο μπορεί να προέλθει από πολλές πηγές

Πηγές Πληροφορίας για το Χρόνο

- Απόλυτες ημερομηνίες και χρονικές στιγμές
  - “on April 6, 1995”
- Σχετικές ημερομηνίες και χρονικές στιγμές
  - “last week”
- Χρόνοι ρημάτων
- Γνώση σχετικά με την έμφυτη αλληλουχία των γεγονότων
- Καθώς η ανάλυση του χρόνου μπορεί να αλληλεπιδρά με άλλα είδη ανάλυσης, συνήθως πραγματοποιείται στο στάδιο συμπερασμού

Αξιολόγηση στα MUCs

- Στους συμμετέχοντες δίνονται αρχικά
  - Μία λεπτομερής περιγραφή του σεναρίου (η πληροφορία που πρέπει να εξαχθεί)
  - Ένα σύνολο κειμένων και των templates που θα εξάγονται από αυτά τα κείμενα (training corpus)
- Οι συμμετέχοντες έχουν κάποιο χρονικό διάστημα στη διάθεσή τους (1-6 μήνες) για να προσαρμόσουν τα συστήματά τους στο νέο σενάριο

Αξιολόγηση στα MUCs

- Μετά από το χρόνο προσαρμογής, κάθε συμμετέχων
  - Λαμβάνει ένα νέο σύνολο κειμένων (test corpus)
  - Χρησιμοποιεί το σύστημά του για να εξάγει πληροφορία από αυτά τα κείμενα
  - Επιστρέφει τα εξαγόμενα templates στον διοργανωτή του συνεδρίου
- Ο διοργανωτής έχει δημιουργήσει χειρονακτικά το σωστό σύνολο των templates από το test corpus
- Σε κάθε σύστημα αποδίδεται μία ποικιλία από scores συγκρίνοντας την απόκρισή του με τις σωστές απαντήσεις

Μετρήσεις Επιδόσεων

- Αν ορίσουμε ως:
  - SLOTS= συνολικός αριθμός slots
  - FILLED = συνολικός αριθμός συμπληρωμένων slots από το σύστημα
  - CORRECT = Συνολικός αριθμός σωστά συμπληρωμένων slots από το σύστημα
- Τότε η επίδοση ενός συστήματος εκφράζεται ως ακρίβεια (precision) και ανάκληση (recall):
  - precision = CORRECT / FILLED
  - recall = CORRECT / SLOTS
- Το F-score είναι ένα μέγεθος που συνδυάζει recall και precision:
  - F = (2 x precision x recall) / (precision + recall)

Υπο-εργασίες (Sub-tasks)

- Named entities recognition (NE)
  - Αναγνώριση των ονοματικών οντοτήτων και ταξινόμησή τους ως άτομα, εταιρείες, οργανισμούς, τοποθεσίες κτλ.
- Identification of template elements (TE)
  - Αναγνώριση των οντοτήτων που καλύπτουν τα template slots
- Recognition of template relations (TR)
  - Αναγνώριση των σχέσεων μεταξύ των οντοτήτων
- Scenario template task (ST)
  - Αναγνώριση των γεγονότων
- Coreference task (CO)
  - Αναγνώριση των οντοτήτων που συν-αναφέρονται

Αποτελέσματα Αξιολόγησης MUC-7