Analyse sémantique automatique

M2 Linguistique Informatique – Paris Diderot

Pascal Amsili / Marie Candito

Résolution d'anaphores

Résolution de coréférence

M Candito

Biblio

- Lappin and Leass, 1994 : An Algorithm for Pronominal Anaphora Resolution. Computational Linguistics, 20(4)
  - cf. résumé ds chapitre Jurafsky et Martin 2008, chapitre 18
- Soon et al, 2001 : A Machine Learning Approach to Coreference Resolution of Noun Phrases. Computational Linguistics, 27(4):521–544.
- Luo et al. 2004 : A mention-synchronous coreference resolution algorithm based on the Bell tree. In ACL 2004
- Denis P. 2007 : New Learning Models for Robust Reference Resolution, Ph.D. dissertation, University of Texas at Austin.

2 Tâches de TAL : RC et RA

- RA : résolution d'anaphores
  - trouver un antécédent pour toute mention anaphorique
    - en général le plus proche
    - en général tâche restreinte aux mentions anaphoriques linguistiquement marquées (pronoms, possessifs etc…)

2 Tâches de TAL : RC et RA

- RC : résolution de coréférence
  - détecter quelles mentions **nominales** sont référentielles
  - et grouper les mentions référentielles référant à la même entité (du monde)
    - = partition

Historique

- comme d'habitude
  - systèmes par règles
    - algo de Hobbes, 78
    - parcours d'un arbre syntaxique, avec match de traits de genre/nb/personne
  - systèmes par heuristiques
    - pondération manuelle de différents paramètres
    - algo célèbre : Lappin et Leass, 94
  - puis par apprentissage (années 90)

RA: algo Lappin & Leass 94

- initialisation : discourse model = liste vide d'entités
  - une entité sera constituée d'un ensemble de mentions
- pour chaque nouvelle phrase
  - diviser par 2 les valeurs de saillance des entités du discourse model
  - pour chaque NP 3eme pers, non pronominal
    - => l'ajouter au discourse model en tant qu'entité singleton
    - et en calculant son score de saillance, d'après les facteurs de saillance "mono" (ne concernant que la mention, et pas un couple mention / pronom)
  - pour chaque NP 3eme pers PROnominal
    - candidats= l'ens. des entités du discourse model, compatibles en genre/nb/pers avec le pronom, et satisfaisant contraintes syntaxiques
    - calculer scores de saillance en incluant les facteurs dépendant du pronom (cataphore, parallélisme)
    - choisir l'entité de score max , la plus proche en cas d'égalité
    - mettre à jour le discourse model :
      - intégrer le pronom à l'entité choisie
      - mettre à jour la valeur de saillance de l'entité
        - travail phrase par phrase
        - pour chaque facteur prendre la valeur max du facteur, pour toutes les mentions de l'entité de la phrase courante

Exemple

- (on applique l'algo également pour les dét. possessifs de 3eme pers)

  Depuis une semaine, Thierry Henry contre-attaque.

  Il s'est longuement exprimé sur RTL et L'Equipe lundi pour répondre à son principal détracteur.

  Depuis quinze jours que [[sa] main] enflamme [la planète football], il n'avait fourni que des explications lapidaires.

- Suggestions d'améliorations?

RA et RC par apprentissage : corpus annotés

- MUC-6 (1995) : 30 doc de train / 30 doc de test
- MUC-7(1998) : 30 doc de train / 20 doc de test
  - annotations XML in situ
  - http://www-nlpir.nist.gov/related_projects/muc/proceedings/co_task.html
  - annotation des mentions coréférentes avec au moins une autre mention
    - pas de singleton
    - coréférence marquée avec la mention précédente la plus proche
- ACE (1999, 2002, ….)
  - annotations XML déportées (offsets)
  - 420 docs de train / environ 100 doc de test (3 domaines)
  - http://www.itl.nist.gov/iad/mig/tests/ace/2004/

- => ces données ont stimulé la recherche en RA et RC
- et mis en évidence des pbs linguistiques
  - relations anaphoriques ne correspondant pas à de la coréférence
    - plutôt entre entités de discours
  - pb des NP attributs => non référentiels
    - => peut créer des incohérences, en particulier si variation ds le temps
    - *Paul, qui était à l'époque le directeur des ventes, est aujourd'hui le président du groupe.*

MUC-6 : exemple

- <s> <COREF ID="0">Ocean Drilling & Exploration Co.</COREF> will sell <COREF ID="3" MIN="business"><COREF ID="2" TYPE="IDENT" REF="0">its</ COREF>contract-drilling business</COREF>, and took a $50.9 million loss from discontinued operations in <COREF ID="12" MIN="quarter">the third quarter</COREF> because of the planned sale. </s>
- <s> <COREF ID="9" TYPE="IDENT" REF="2" MIN="company">The New Orleans oil and gas exploration and diving operations company</ COREF> added that <COREF ID="10" TYPE="IDENT" REF="9">it</ COREF> doesn't expect any further adverse financial impact from the restructuring. </s>
- ...
- <s> In <COREF ID="11" TYPE="IDENT" REF="12" MIN="quarter">the third quarter</COREF>, <COREF ID="13" TYPE="IDENT" REF="10" MIN="company">the company, which is 61%-owned by Murphy Oil Corp. of Arkansas,</COREF> had <COREF ID="100" MIN="loss">a net loss of <COREF ID="17" TYPE="IDENT" REF="100">$46.9 million</COREF>, or <COREF ID="16" TYPE="IDENT" REF="17" MIN="91 cents">91 cents a share</COREF>. </s>
- ...

Corpus pour le français

- Tutin et al. 2000 : 1 M de mots
  - expressions anaphoriques grammaticales
- dede (descriptions définies)
  - Gardent & Manuelian 2006
- ANCOR (ANaphore et Coréférence dans les Corpus ORaux)
  - Lefeuvre et al. 2014

RA : par apprentissage supervisé

- souvent restriction à des types d'anaphores
  - par exemple : les pronoms / les possessifs
- construction classifieur binaire
  - objet à classer : couple (anaph, candidat-antécédent)
  - 2 classes : le couple est coref / le couple n'est pas coref
  - **sampling** (=construction exemples d'apprentissage)
    - doit être lié à l'algo de choix de l'antécédent, lors de la phase de prédiction
    - sampling exhaustif McCarthy and Lehnert (1995) :
      - exemples positifs = tous les couples [x / mention antérieure coréférente à x]
      - exemples négatifs = tous les couples [x / mention antérieure NON coréférente à x]
      - => apprentissage long, cf. bcp d'exemples au sein d'un doc
    - réduction du nb d'exemples : par ex (Soon et al. 2001):
      - exemple positif : le premier candidat coréférent avec l'anaph, à gauche de l'anaph
      - exemples négatifs : les candidats entre l'antécédent et l'anaph
    - il existe aussi : sampling variable selon la catégorie de l'anaphore
      - en particulier, on peut imposer : antécédent d'une anaphore non pronominale ≠ pronom

RA : par apprentissage supervisé

- Phase de prédiction :
  - phase 1 optionnelle : identification des anaphores
    - par ex. filtrage des explétifs
  - phase 2 :
    - pour une anaphore A : récupérer les candidats antécédents dans la phrase courante et les x phrases précédentes
      - typiquement : tous les NPs (y compris les pronoms) , x = 1 ou 2
      - doit être cohérent avec le sampling réalisé pour l'apprentissage
    - leur appliquer le classifieur
    - sélectionner une réponse parmi les candidats, d'après le classifieur
    - plusieurs stratégies pour cela :
      - closest-first : parmi les candidats classés comme coréférents à *A*, choisir le plus proche à gauche de l'anaph
      - best-first : parmi les candidats, choisir celui de score maximal d'après le classifieur
      - idem en cohérence avec le sampling

RA supervisé variantes

- spécialiser le classifieur binaire
  - en particulier : un classifieur par catégorie d'anaphores
    - cf. contraintes très différentes entre par ex. pronoms et noms propres
      - pronoms : ont leur antécédent dans les 1 ou 2 phrases précédentes max
      - nom propres (par ex. formes abrégées de nom propre) : potentiellement tout le document
- apprentissage d'un ranker au lieu d'un classifieur

  (Denis, 2007)
  - le but est de trouver le meilleur antécédent possible, et pas de trouver tous les antécédents possibles
  - => ranking : intégration de cette contrainte dès l'apprentissage, et pas uniquement dans l'algo de prédiction

Parenthèse : ranking

- ranking : adapté pour une tâche où l'on doit choisir une solution parmi un ensemble de candidats
- exemples
  - choisir un arbre syntaxique parmi les n-best analyses fournies par un parser ("reranking", Johnson & Charniak, 05)
  - RA : choisir un antécédent parmi un ens. d'antécédents possibles

Parenthèse : ranking perceptron

- Si on utilise un classifieur pour la tâche de TA, on utilise un classifieur **binaire**

- Rappel : apprentissage perceptron binaire :
- on dispose de T exemples ($x^{(t)}$, $y^{(t)}$)
  - avec $x^{(t)}$ = la représentation vectorielle d'un couple anaphore + candidat-antécédent
  - $y^{(t)}$ = 0 ou 1 (coréférence / non coréférence)
- pour i de 1 à ITER
  - pour t de 1 à T
    - si $x^{(t)}$ n'est pas bien classé lorsque l'on utilise le vecteur w courant
      - c'est-à-dire si $y^{(t)}_{pred}$ = signe(w.$x^{(t)}$) ne vaut pas $y^{(t)}$
    - alors : mettre à jour w
      - w := w + $y^{(t)}x^{(t)}$

Parenthèse : ranking perceptron

- Mais en ranking : tout se passe comme si les candidats antécédents étaient les classes
- => on doit donc utiliser un classifieur **multiclasse**

- Rappel apprentissage perceptron multiclasse :
  - on considère C classes $y_1,\ldots,y_C$
  - on dispose de T exemples :
    - notés $(x^{(t)}, y^{(t)})$ avec $y^{(t)} \in \{y_1,\ldots,y_C\}$
  - pour i de 1 à ITER
    - pour t de 1 à T
      - $y_{pred} = \text{argmax}_{y \text{ in } y1\ldots yC}\ w.f(x^{(t)}, y)$
      - si $y_{pred} \neq y^{(t)}$
        - update : $w = w + f(x^{(t)}, y^{(t)}) - f(x^{(t)}, y_{pred})$

Parenthèse : ranking perceptron (Collins, 2002)

- exactement comme un perceptron multiclasse
- mais tout se passe comme si les "classes" étaient les candidats de chaque exemple
- apprentissage :
  - on dispose de T exemples ($x^{(t)}$, candidats($x^{(t)}$))
  - avec $c^{(t)*}$ = le candidat gold
  - pour i de 1 à ITER
    - pour t de 1 à T
      - $c_{pred} = \text{argmax}_{c \text{ in cand}(x(t))} w.f(x^{(t)}, c)$
      - si $c_{pred} \neq c^{(t)*}$
        - update : $w = w + f(x^{(t)}, c^{(t)*}) - f(x^{(t)}, c_{pred})$

Parenthèse : ranking perceptron

(Collins, 2002)

- la phase de prédiction est exactement comme le cas "best-first" vu précédemment :
  - on choisit le candidat-antécédent de score max d'après le ranker
- la différence intervient uniquement à l'apprentissage :
  - au lieu d'avoir un exemple par couple [anaphore / candidat-antécédent]
  - on a un exemple par anaphore
    - => et on intègre dès l'apprentissage la contrainte qu'un et un seul candidat-antécédent va être choisi
    - Voir par exemple la différence à l'apprentissage dans le cas de 2 candidats de score proches

CR par apprentissage

- Fondé sur
  - apprentissage de classifieurs (ou rankers) pour la tâche RA
  - construction des chaînes de coréf (ou "**entités**")
    - l'ensemble des entités correspond formellement à une partition des mentions
      - tâche de clustering
      - avec éventuellement des entités singletons (mentions non anaphoriques, sans reprise ultérieure par une anaphore)
    - Rem : le nombre de partitions possibles est énorme
      - nombre de partitions d'un ensemble à n éléments (ici n mentions) = nombre de Bell
        - par ex. pour n=20 => de l'ordre de 10^{23}

CR par apprentissage

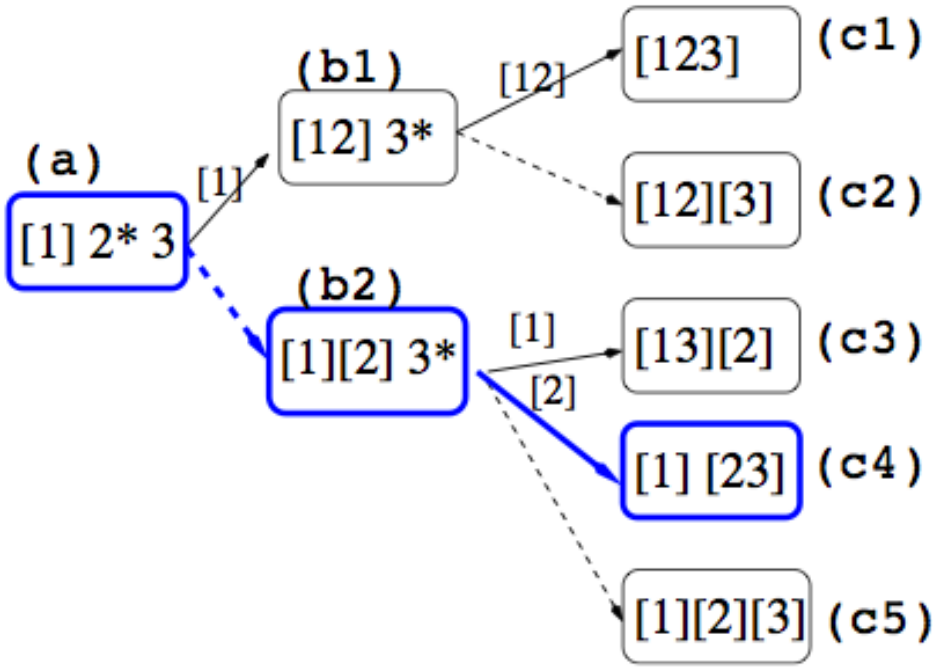
- Système de base :
  - **Apprentissage** :
    - Apprentissage de classifieur pour tâche RA
    - pour tout type de mentions anaphoriques (pas uniquement linguistiquement marquées, i.e. pronoms)
    - éventuellement en utilisant en amont un classifieur dédié pour prédire si une mention est "référentielle" ou pas
  - **Prédiction** :
    - Pour chaque mention (référentielle) : choix du meilleur ou du plus proche antécédent
    - puis **constitution des chaînes de coréférence**
      - basique : dans cas où le RA fournit au plus 1 antécédent par mention => la **clôture transitive** de la relation de coréférence induit une partition des mentions => i.e. les chaînes de coréférence

CR par apprentissage

- Enjeu : prise de décision moins locale
  - => i.e. constituer les entités en faisant intervenir plusieurs antécédents pour une mention
    - intuitivement plus riche : le but est de maximiser la cohérence au sein d'une entité
      - exemple : ***Mrs Clinton*** *declared that* ***Barack Obama*** *blabla.* ***Clinton*** *then said that* ***Obama*** *blabla.* ***She*** *…*
  - nécessite de considérer plusieurs antécédents pour une mention :
    - MAIS : incohérences potentielles à résoudre
      - par ex. si prédiction coref(A,B) et coref(B,C) et not(coref(A,C))
- champ de recherche très actif…

CR : variantes : par ex. Luo et al. 04

- mentions par ordre linéaire $m_1 m_2 \ldots m_n$
- l'ensemble des partitions des mentions est représenté sous forme d'un arbre de Bell
  - racine = m1, puis une profondeur par mention
  - 2 types d'arc lors du passage à une mention supplémentaire m_i:
    - (pointillés) lien de type "start" = création d'une nouvelle entité singleton { m_i }
    - lien de type "link" = ajout de m_i à une entité existante



CR : variantes : par ex. Luo et al. 04

- modèle probabiliste affectant un score à chaque arc de l'arbre
  - P(arc link entre mention_i et entité_j | entités courantes, mention_i, entité_j)
  - approximée par $\max_{\text{mention\_k}}$ P(coref entre mention_i et mention_k)
    - on retrouve un score de classification de type RA entre 2 mentions
    - mais max sur toutes les mentions de l'entité
- algorithme de programmation dynamique pour construire efficacement l'ensemble des n-meilleurs chemins
  - = n meilleures partitions
  - = **beam search** (recherche par faisceau)
  - cf. recherche exacte du meilleur chemin : inaccessible computationnellement
- crucialement on obtient un modèle plus global car:
  - score global sur toute la partition
  - conservation des n meilleurs chemins
    - au final on peut choisir un chemin passant par un arc non optimal localement

CR : variantes

- Denis, 2007 : intégration de contraintes globales lors de la prédiction des chaînes de coréf, via ILP
  - prise en compte de tous les scores de coref entre 2 mentions
  - capture via contraintes ILP
    - de contraintes de transitivité :
      - si coref(A,B) et coref(B,C) alors on devrait avoir coref(A,C)
    - d'un classifieur externe qui prédit le type d'une mention (personne, loc, organisation etc…)
      - => contraintes ILP pour imposer l'uniformité de type au sein d'une entité
    - d'un classifieur externe qui prédit le caractère "déjà connu" versus "nouveau ds le discours" d'une mention

ILP

- *Integer Linear Programming*
  - ou *Programmation linéaire en nombres entiers*
  - ou encore *Optimisation linéaire en nombres entiers*

- = Cas particulier de la programmation linéaire
  - découverte fin années 30 / années 40
  - par les mathématiciens Kantorovitch (russe), Dantzig, Von Neumann (américains)
- = problème d'optimisation :
  - apprentissage des x_i maximisant ou minimisant une fonction objective linéaire de variables x_i
  - $c_1x_1+c_2x_2 + \dots c_nx_n$
  - sous contraintes également linéaires, de type inégalités
    - $a_1x_1+b_2x_2 + \dots b_nx_n < a$
    - $b_1x_1+b_2x_2 + \dots b_nx_n < b \dots$

- en ILP : les x_i sont des entiers (cf. par ex. déf de Wolsey, 88)
  - en outre, les x_i sont souvent restreints à prendre les valeurs 0 ou 1
    - en particulier lorsqu'une variable x_i correspond à une décision binaire
  - ce qui donne une complexité de type NP-difficile
    - mais algos existent, souvent efficaces en pratique (pour qqs dizaines / centaines de variables)
    - des solveurs sont disponibles CPLEX, LP_SOLVE

ILP : exemple

maximize: $x_1 + x_2$

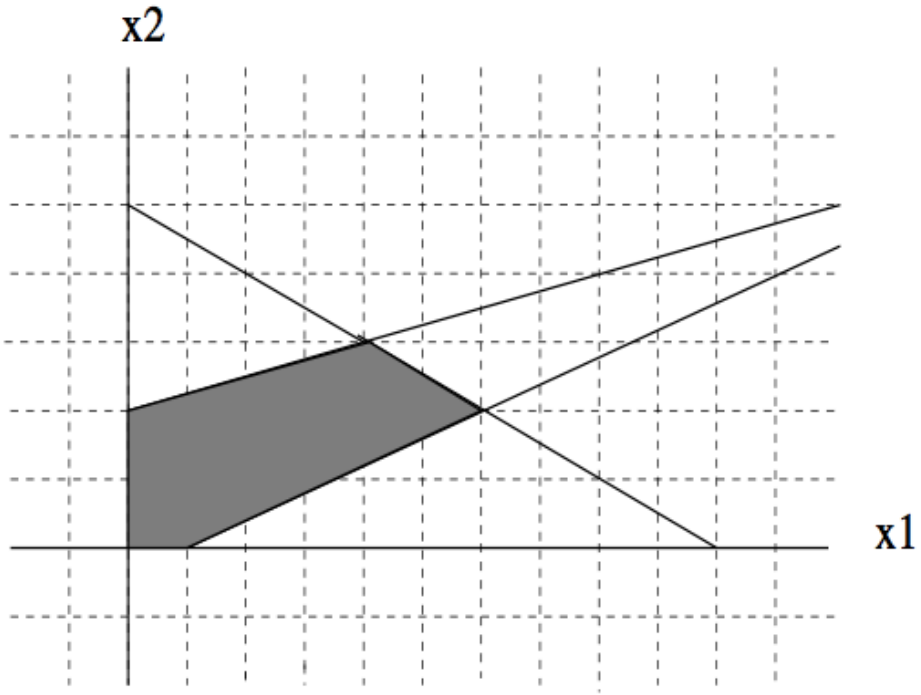
subject to: $x_1, x_2 \geq 0$

$x_1 + 2x_2 \leq 10$

$4x_2 - x_1 \leq 8$

$2x_1 - 5x_2 \leq -2$

solutions: $x_1 = 6; x_2 = 2$



RC via ILP (Denis, 07)

- variables x_{ij} = 1 si m_i et m_j coréfèrent, 0 sinon
- on dispose d'un classifieur RA de type Maxent
- on note
  - p_{ij} = P(coref | m_i, m_j)
  - q_{ij} = P(non coref | m_i,m_j) = 1- P(coref | m_i, m_j)
  - (sont donc entre 0 et 1)
- Objectif :
  - choisir les meilleures décisions de coréférence binaire
  - i.e. choisir x_{ij} = 0 ou 1
  - d'après les probas de RA (les p_{ij} et q_{ij})
  - en intégrant des contraintes globales via des inégalités ILP
  - puis fusion "agressive" : une mention est groupée avec tous ses coréférents précédents

RC via ILP (Denis, 07)

- modèle de base :
- => formulation ILP pour la tâche de base : trouver les x_{ij} *minimisant*

$$\sum_{i,j} -\log(p_{ij})x_{ij} - \log(1-p_{ij})(1-x_{ij})$$

- i.e. poids c_{ij} fixés à $-\log(p_{ij}) + \log(1-p_{ij})$ (+constante)
  - si p_{ij} proche de 1 : on a intérêt à ce que x_{ij} vaille 1
  - si p_{ij} proche de 0 : on a intérêt à ce que x_{ij} vaille 0
- équivaut en fait à choisir $x_{ij} = 1$ ssi $p_{ij} > 0.5$
- puis fusion

RC via ILP (suite)

- intégration d'un classifieur prédisant le "statut discursif" d'une mention : "déjà connu" versus "nouveau ds le discours"
- => en ILP :
  - variables y_i vaut
    - 1 si m_i "déjà connu dans le discours"
    - et 0 sinon (et donc à ne pas relier à un antécédent précédent)
  - scores du classifieur binaire déjà connu / nouveau
    - on note o_i = P(déjà connu | m_i)
  - fonction à minimiser :

$$\sum_i -\log(o_i)y_i - \log(1-o_i)(1-y_i) + \sum_{i,j} -\log(p_{ij})x_{ij} - \log(1-p_{ij})(1-x_{ij})$$

  - => en l'absence d'inégalités, revient à choisir
    - x_{ij} =1 ssi p_{ij} > 0.5
    - y_i = 1 ssi o_i > 0.5

RC via ILP (suite)

- Formulation de contraintes pour capturer la **cohérence** entre les décisions "connu" / "non connu" et les décisions "coréférent" / "non coréférent"
  - contrainte "**résoudre toutes les anaphores**" : si m_i est "déjà connu" => il doit coréférer avec au moins un m_j
    - => au moins un x_{ij} doit valoir 1
  - contrainte "**ne résoudre que les anaphores**" : si m_i coréfère avec un m_j (j > i) alors m_j doit être "déjà connu"
    - => si x_{ij} vaut 1 alors y_j doit valoir 1
  - formulables par les inégalités suivantes :

$$\forall i \quad y_i \leq \sum_j x_{ij} \quad \text{et} \quad \forall i < j \quad y_j \geq x_{ij}$$

RC via ILP (suite)

- Intégration de contraintes de transitivité
  - on peut formuler sous forme d'inégalités les contraintes :
    - $x_{ij}=x_{jk}=1 \Rightarrow x_{ik}=1$ transitivité

$$\forall i,j,k \;\; x_{ij} + x_{jk} \leq x_{ik} + 1$$

    - $x_{ij}=x_{ik}=1 \Rightarrow x_{jk}=1$ euclidianité

    - $x_{ij}=x_{kj}=1 \Rightarrow x_{ik}=1$ anti-euclidianité

RC via ILP (suite)

- Intégration de contraintes sur les types des entités :
  - on considère un ensemble de types $T_1 \ldots T_T$ de types d'entités nommées (person, organization…)
  - et on dispose d'un classifieur probabiliste permettant de scorer l'assignation d'un type à une mention $P(T_t \mid m_i)$
- question : comment intégrer cela au résolveur de coréférence ILP vu précédemmment?

RC via ILP (suite)

- Intégration de contraintes sur les types des entités :
  - ILP => variables z_{it} pour tout couple mention m_i / type T_t
    - valant 1 si m_i a le type T_t et 0 sinon
    - on note $r_{it} = P(T_t \mid m_i)$
  - idem on ajoute à la fonction objective à minimiser : $-\log(r_{it})z_{it} + -\log(1-r_{it})(1-z_{it})$
  - contraintes
    - au plus un type par mention
    - cohérence entre typage et coréférence : si x_{ij} vaut 1 alors tous les $z_{it} - z_{jt}$ valent 0
    - formulable (entre autres) par :

$$\forall i \quad 1 \leq \sum_t z_{it} \leq 1$$

$$\text{et} \quad \forall i,j,t \quad 1 - x_{ij} \geq z_{it} - z_{jt} \text{ et } 1 - x_{ij} \geq z_{jt} - z_{it}$$

RC via ILP : résultats

- les résultats montrent que le modèle joint combinant
  - coref
  - structure informationnelle
  - type d'entité nommées
  - anti-euclidianité
  - améliore toutes métriques confondues par rapport à la classification binaire avec construction "best-first" des entités