Résultats et méthodes de détection de co-références déployées lors de la campagne d'évaluation CoNLL 2011.

Présenté par *Eric Charton*

ÉCOLE POLYTECHNIQUE MONTRÉAL

Recherche menée par Michel Gagnon et Eric Charton

Plan

Méthodes et principes de la détection de co-références

Qu'est ce qu'une chaine de co-références?

Texte:

Outre sa descendance directe, c'est désormais
à ses petits fils que Kim-Jong-Il va devoir rappeller
quelques principes de base. Le petit dernier, Kim Han Sol,
s'est fait en quelques jours une réputation sur les
réseaux sociaux. La descendance du dictateur lui cause
du soucis.

Qu'est ce qu'une chaine de co-références?

Mentions:

Outre sa descendance directe, c'est désormais à ses petits fils que Kim-Jong-Il va devoir rappeller quelques principes de base. Le petit dernier, Kim Han Sol, s'est fait en quelques jours une réputation sur les réseaux sociaux. La descendance du dictateur lui cause du soucis.

Qu'est ce qu'une chaine de co-références?

Chaines:

Outre sa descendance directe, c'est désormais à ses petits fils que Kim-Jong-Il va devoir rappeller quelques principes de base. Le petit dernier, Kim Han Sol, s'est fait en quelques jours une réputation sur les réseaux sociaux. La descendance du dictateur lui cause du soucis.

Qu'est ce qu'une chaine de co-références?

Imbrications:

Outre sa descendance directe, c'est désormais à ses petits fils que Kim-Jong-Il va devoir rappeller quelques principes de base. Le petit dernier, Kim Han Sol, s'est fait en quelques jours une réputation sur les réseaux sociaux. La descendance du dictateur lui cause du soucis.

Nature des objets textuels co-référents

- Pronoms: il, elle, nous
- Entités nommées: Montréal, Jean Dutour, École Polytechnique de Montréal, I-Pad, F18
- Dates: 21 juin 2011, équinoxe de printemps, hier, demain, 12h, jour de l'an
- Syntagmes nominaux: le fils du président, l'extraordinaire avion de chasse, la tablette Apple
- Des ensembles: les fille du chef d'état, Marie et ses soeurs, les habitants, la population, la famille princière

Principes de détection

Exploiter plusieurs niveaux de connaissances dans des étapes successives.

Une succession de tâches

1. **Contenus**: déterminer la nature des objets textuels et la structure du texte (POS, EN, Syntaxe...).
2. **Mentions**: identifier les objets textuels susceptibles d'être co-référents.
3. **Relations**: déterminer les relations de co-références entre objets textuels.

Les campagnes d'évaluation et leurs métriques

Métriques et évaluation

Problème ouvert: comment mesurer l'imprécision qui n'est pas une erreur ?

Problématique

- Évaluer à la fois un étiquetage (les mentions) et des relations entre mentions (les chaînes de co-références).
- Une chaîne de co-référence unique identifiée en deux groupes par un système est partiellement valable.
- Les deux aspects (mentions et relations) sont dépendants.
- Comment considérer les mentions incomplètes, surnuméraires, absentes?

Métriques et évaluation

Métrique MUC-6-7: compte les liens entre mentions d'après un Gold Standard.

- La chaîne C_i (les clefs) de mentions parfaite $m1, ..., m_n$.
- La chaine K_i fournie par le système, considérée équivalente à C_i.
- On ne conserve dans K_i que les mentions identiques à celles contenues dans C_i.
- Précision et rappel
- F-Score classique

$$recall = \frac{\sum_{1..i}(|C_i| - |K_i|)}{\sum_{1..i}(|C_i| - 1)} \quad precision = \frac{\sum_{1..i}(|K_i| - |C_i|)}{\sum_{1..i}(|K_i| - 1)}$$

Défaut de MUC-6-7: peut fiable si beaucoup de références sont des singletons de mentions, ou si des chaines sont divisées.

Mode de validation des K_i très variable (exemple: prise en compte du Span dans CoNLL 2011).

Métriques et évaluation

Propositions alternatives:

Autres métriques

- B-CUBED : complète MUC; calcule P et R pour toutes les mentions du documents puis les combines dans P et R final (plusieurs variantes).
- CEAF : critique de B3 qui utilise les mentions plus d'une fois. Aligne les mentions et les clés du Gold Standard pour le calcul (plusieurs variantes).
- BLANC: le plus récent (2010). Utilise l'index RAND pour mieux évaluer les chaînes par leurs regroupements.

MUC basé sur le lien, B3 basé sur les mentions, CEAF basé sur les groupes. Les évaluations récentes proposent donc un score global équivalent à la somme de n métriques divisé par n.

CoNLL 2011: (MUC + B3 + CEAF) / 3

Bonne revue de détails dans *Evaluation Metrics For End-to-End Coreference Resolution Systems*, SIGDIAL 2010.

Évaluations précédentes

MUC-6 1995 - MUC-7 1997 - ACE 2005

- Tache proche de CoNLL 2011. Texte variés et livrés avec layers (EN, NP ...)
- Identifier les mentions, déterminer les chaines de co-références
- Métrique MUC ou B3. Seulement 60 k mots.

Évaluations précédentes

MUC-6 1995 - MUC-7 1997 - ACE 2005

- Tache proche de CoNLL 2011. Texte variés et livrés avec layers (EN, NP ...)
- Identifier les mentions, déterminer les chaines de co-références
- Métrique MUC ou B3. Seulement 60 k mots.

INLG 2010 - GREC Challenge

- Texte brut mais simple: localiser uniquement les EN de personnes dans des fiches biographiques Wikipédia
- Identifier les mentions, déterminer les chaines de co-références
- Métriques fusionnées (MUC, B3, Blanc)

Évaluations précédentes

MUC-6 1995 - MUC-7 1997 - ACE 2005

- Tache proche de CoNLL 2011. Texte variés et livrés avec layers (EN, NP ...)
- Identifier les mentions, déterminer les chaines de co-références
- Métrique MUC ou B3. Seulement 60 k mots.

INLG 2010 - GREC Challenge

- Texte brut mais simple: localiser uniquement les EN de personnes dans des fiches biographiques Wikipédia
- Identifier les mentions, déterminer les chaines de co-références
- Métriques fusionnées (MUC, B3, Blanc)

Semeval 2010: Coreference Resolution in Multiple Languages

- Texte variés (nouvelles, transcriptions ...). Les mentions sont identifiées (corpus Ontonotes 2.0-BBN)
- Les singletons sont comptés
- Seule la détection de chaînes est mesurée

Évaluations précédentes

Avant CoNLL 2011,les résultats sont difficilement interprétables

- INLG ne concerne que les personnes. Corpus faciles (biographies Wikipédia).
- Semeval ne présente pas la totalité de la tâche.
- Le corpus MUC est trop ancien et petit.
- Les systèmes déployés lors des campagnes MUC sont obsolètes.

CoNLL 2011 Shared Task

12 ème année depuis 2000. Particularités:

- Corpus Ontonotes: 1,3 millions de mots en anglais, richement annoté (syntaxe, propositions, sens, entités nommées, événements, co-références).
- Genres de documents multiples: transcriptions, articles, conversations, blogs ...
- Très haute qualité de l'annotation manuelle, vérifiée par les accords inter-annotateurs.
- Pas de singletons dans les co-références: seules les chaines réelles sont annotées
- Deux familles de co-références: identité et appositions.

Ontonotes

```
#begin document (nw/wsj/07/wsj_0771); part 000
...
...
nw/wsj/07/wsj_0771 0   0        ``   ``  (TOP(S(S*     -    -  -   -        *       *     (ARG1*  *       -
nw/wsj/07/wsj_0771 0   1 Vandenberg  NNP       (NP*     -    -  -   -  (PERSON)  (ARG1*        *  *  (8|(0)
nw/wsj/07/wsj_0771 0   2       and   CC          *     -    -  -   -        *       *        *  *       -
nw/wsj/07/wsj_0771 0   3   Rayburn  NNP         *)     -    -  -   -  (PERSON)      *)        *  * (23)|8)
nw/wsj/07/wsj_0771 0   4       are  VBP       (VP*    be   01  1   -        *    (V*)        *  *       -
nw/wsj/07/wsj_0771 0   5    heroes  NNS   (NP(NP*)     -    -  -   -        *  (ARG2*        *  *       -
nw/wsj/07/wsj_0771 0   6        of   IN       (PP*     -    -  -   -        *       *        *  *       -
nw/wsj/07/wsj_0771 0   7      mine   NN   (NP*))))     -    -  5   -        *      *)        *  *    (15)
nw/wsj/07/wsj_0771 0   8         ,    ,          *     -    -  -   -        *       *        *  *       -
nw/wsj/07/wsj_0771 0   9        ''   ''         *)     -    -  -   -        *       *       *)  *       -
nw/wsj/07/wsj_0771 0  10       Mr.  NNP       (NP*     -    -  -   -        *       *     (ARG0*  *     (15
nw/wsj/07/wsj_0771 0  11     Boren  NNP         *)     -    -  -   -  (PERSON)       *       *)  *     15)
nw/wsj/07/wsj_0771 0  12      says  VBZ       (VP*   say   01  1   -        *       *      (V*)  *       -

#end document
```

Architecture d'un système de détection et de résolution de co-références

Détection des mentions

Utiliser les diverses informations fournies par les étiquettes du corpus pour localiser les mentions candidates:

Type d'informations utilisées

- Les entités nommées sont localisées par leur étiquette.
- Les syntagmes nominaux par la syntaxe (NP) et le DT.
- Les dates et heures par des séquences de mots (*tonight*) ou des étiquettes d'EN.
- Les pronoms par l'étiquette morphosyntaxique et le mot.

Les erreurs potentielles sont filtrées par des règles (exemple *It is*). Lorsque le corpus est uniquement textuel (ex GREC-INLG 2010), il doit être étiquetée préalablement.

Identification des mentions co-référentes par apprentissage sur des paires [Soon et al 2001]

Considérant une suite S de mentions $m_{0...z}$:

Deux mentions de S sont elles co-référentes ?

- Considérer toutes les mentions contenues dans un intervalle I sur S.
- Intégrer toutes les combinaisons de paires de mentions successives sur l'intervalle I avec leurs caractéristiques (distance, nature, etc) dans des vecteurs.
- **Apprentissage**: un vecteur est étiqueté 1 s'il contient une paire co-référente.
- **Détection**: soumettre un vecteur au classifieur qui répond: 0/1.

La réunion des paires de mentions co-référentes au sein d'une chaîne intervient à postériori.

Principe d'apprentissage

The FBI says its New York office will lead the agency's investigation into the Cole Bombing.
It's the same office that's investigated similar terrorist attacks [...]

Mention detection

Mention / Coreference / Coreference label number / cluster

1	[A]	The FBI
2	[A]	its
3	[E]	its New York office
4	[C]	New York
5	[A]	the agency's
6	[F]	the agency's investigation into the Cole Bombing
7	[B]	the Cole bombing
8	[D]	Cole
9	[E]	It's

Features vectors construction

A (5) (4) (3) (2) -> {m5,m4}{m5,m3}{m5,m2}

A (2) (1) -> {m2,m1}

E (9) (8) (7) (6) (5) (4) (3)
-> {m9,m8}{m9,m7}...{m9,m3}

Caractéristiques du vecteur

Plus le vecteur contient de caractéristiques évoluées, et plus son caractère discriminant sera pertinent pour identifier les paires co-référentes:

Exemples

- Genre (Un féminin ne peut co-référer avec un masculin)
- Nombre (un pluriel ne peut co-référer avec un singulier)
- Distance (la probabilité de co-référer décroit avec la distance - principe de proéminence-saillance)
- Nature (une EN de type personne ne peut co-référer avec une EN de type ORG)

C'est le choix des paramètres du vecteur qui conditionne les performances du modèle de Soon.

Classifieurs

Tout classifieur susceptible de donner une réponse binaire ou probabiliste pour déterminer la classe d'un vecteur est valable.

Exemples

- Perceptron
- SVM
- Arbres
- ...

Score	Mentions			B3			CEAF			MUC		
	R	P	F	R	P	F	R	P	F	R	P	F
MLP	65.91	64.84	65.37	66.61	62.09	64.27	50.18	50.18	50.18	54.47	50.86	**52.60**
SVM	65.06	66.11	**65.58**	65.28	57.68	61.24	46.31	46.31	46.31	53.30	50.00	51.60
J48	66.06	64.57	65.31	66.53	62.27	**64.33**	50.59	50.59	**50.59**	54.24	50.60	52.36

Assemblages des paires dans des chaînes de co-références

Méthodes et heuristiques variées

- Parcours de tableaux pour relier les paires entre elles
- Dépilage par cluster
- Rattachement de clusters par des heuristiques (Ex: entités nommées similaires, identicité, impossibilité logique)
- Accès à des ressources externes (ex: étiquetage sémantique, mais interdit à CoNLL)

Question théorique ouverte: quelle est l'influence réelle de ces heuristiques sur le résultat final ?

Choix des méthodes

Systèmes dominants et points communs:

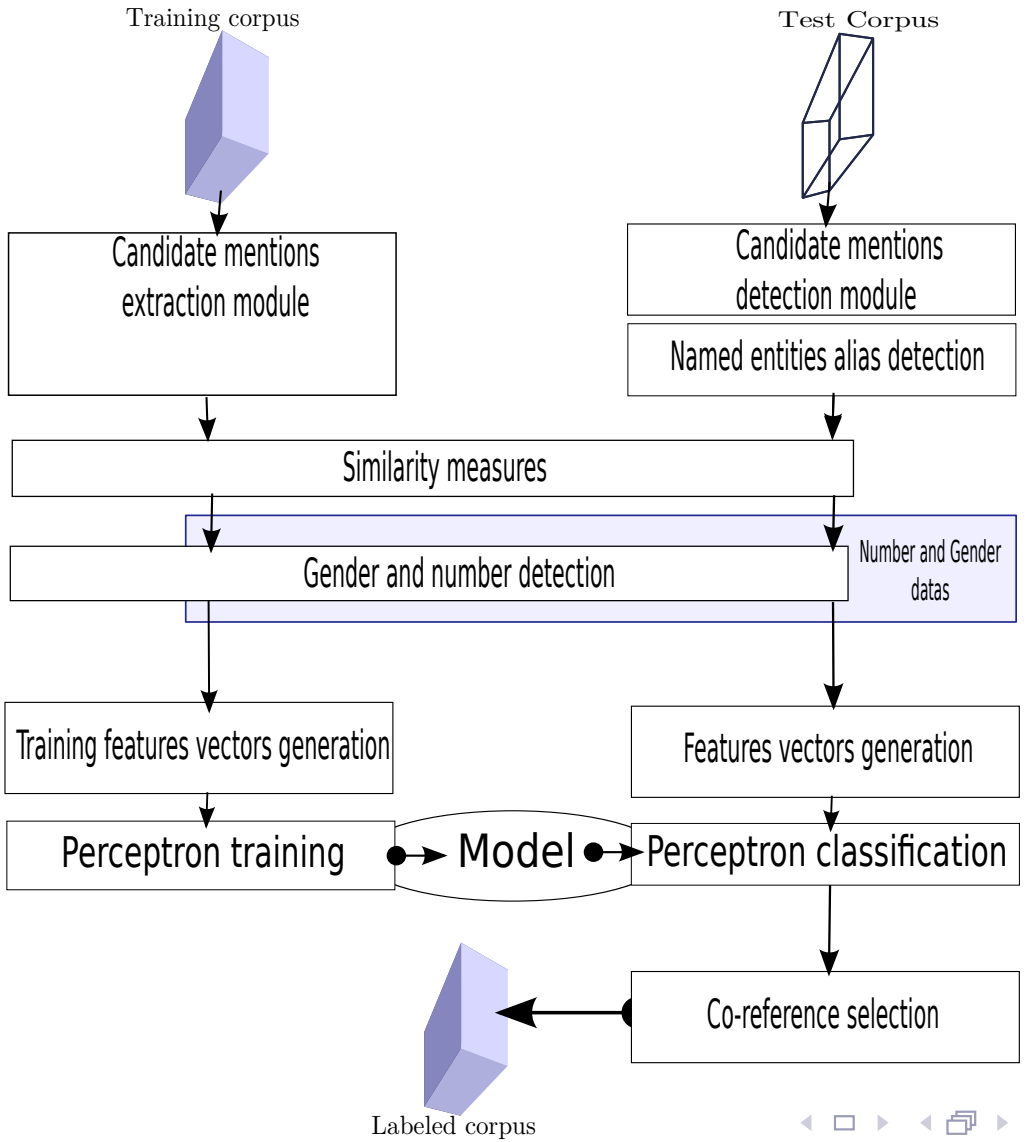
- La plupart des systèmes séparent reconnaissance des mentions et des liens en deux tâches.
- La plupart des systèmes s'inspirent du modèle de Soon de classifieurs de paires.
- La sélection des features est le principal point de recherche.

Alternatives n'utilisant pas la classification de paires:

- Méthode de partitionnement de graphe (Sapena 2010).
- Méthode de détection par MLN (Poon et Domingos 2008).
- Méthodes à base de règles ou hybridées ...

Architecture du système Poly-co

Architecture de Poly-co 2



Détection de caractéristiques de mentions

Utiliser des caractéristiques de mentions les plus riches possibles.

Exemples

- Détection du genre (neutre, féminin, masculin)
- Détection du nombre (pluriel, masculin)
- Normalisation d'une date (ex: hier=j-1 ou demain=j+1)
- Mesurer la similarité de syntagmes par le lexique (ex: Le beau navire azur=Le grand navire bleu)
- Déterminer la similarité des entités nommées (ex: Paris=Ville Lumière)
- Identifier les Alias (ex *Steve Jobs*=*S. Jobs*=*Mr Jobs*=*Steve*)

Détection de caractéristiques de mentions

Intégrer la nature sémantique de certaines mentions dans les caractéristiques du vecteur.

Exemples

- NE SEMANTIC TYPE un des 18 types d'EN (PERSON, ORG, TIME, etc)
- PRP NAME 30 valeurs de pronoms personnels (he, she, it, etc).
- NP NAME valeur indiquant le DT d'une NP (the, this, these, etc).
- NP TYPE indique si NP est démonstratif, défini, ou un quantifieur.

Caractéristiques du vecteur de Poly-co 2

Features names	Parameters	Values
-relation features	m_a and m_b	
IsAlias	yes/no	1/0
IsSimilar	real	0.00 /1.00
Distance	int	0/const(b)
Sent	int	0/x
-specific features	Mention m_a	
IsNE	yes/no	1/0
IsPRP	yes/no	1/0
IsNP	yes/no	1/0
NE_Semantic type	null / EN	0 / 1-18
PRP_name	null / PRP	0 / 1-30
NP_name	null / DT	0 / 1-15
NP_Type	null / TYPE	0 / 1-3
Gender	M/F/N/U	1/2/3/0
Number	S/P/U	1/2/0
Mention m_b Same features as m_a		

Règles de filtrage des chaînes

Le processus de filtrage exclut certaines mentions avant de les soumettre dans des paires au classifieur:

Exemples de règles d'exclusion

- Un pronom ne peut être suivi immédiatement par le verbe *to be* et un pronom relatif dans les 6 mots suivants.
- recherche de motif *it ... to*.
- recherche de motif *it ... be because*.
- recherche de motif *all of us, all of you*.

Assemblages des mentions au sein des chaînes

A l'issue du processus de classification, un tableau indique pour chaque mention son éventuel antécédent co-référent.

L'assemblages des mentions au sein des chaînes de co-références est ensuite réalisée via des processus successifs:

1. Mise en relation des paires.
2. Extraction par dépilage successifs dans des clusters.
3. Fusion éventuelle des clusters.
4. Retrait des mentions non reliées (singletons).
5. Renumérotation.

Résultats de la campagne et commentaires

Résultats généraux

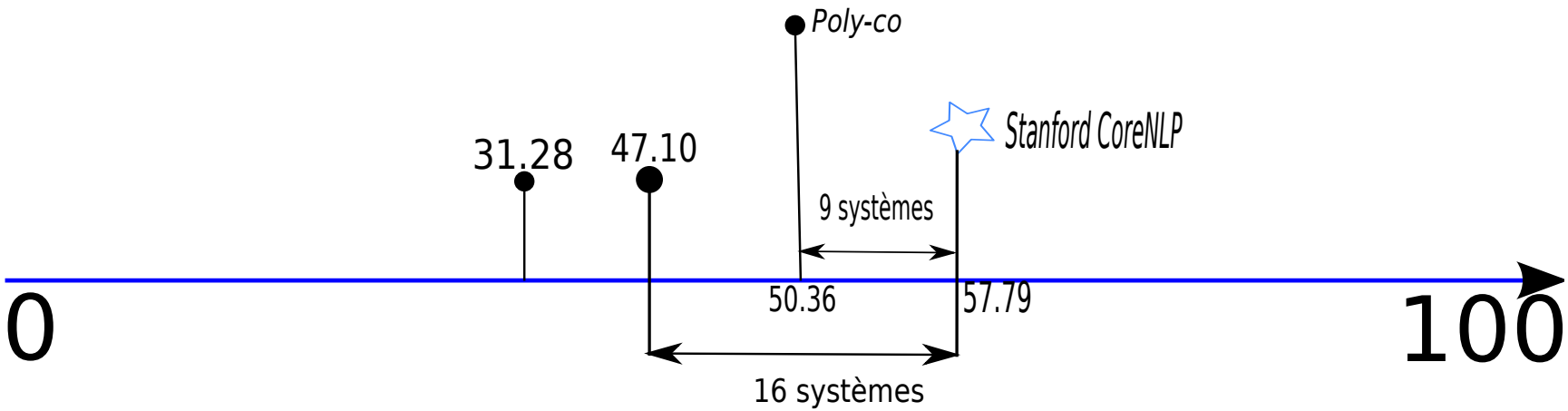
Le premier système utilise des règles appliquées sur des mentions. Les dix systèmes suivants:

- Utilisent l'architecture de Soon pour les features (profondeur variable de 3 à 10 mentions)
- Adoptent des méthodes de classification très variés

Classifieurs:

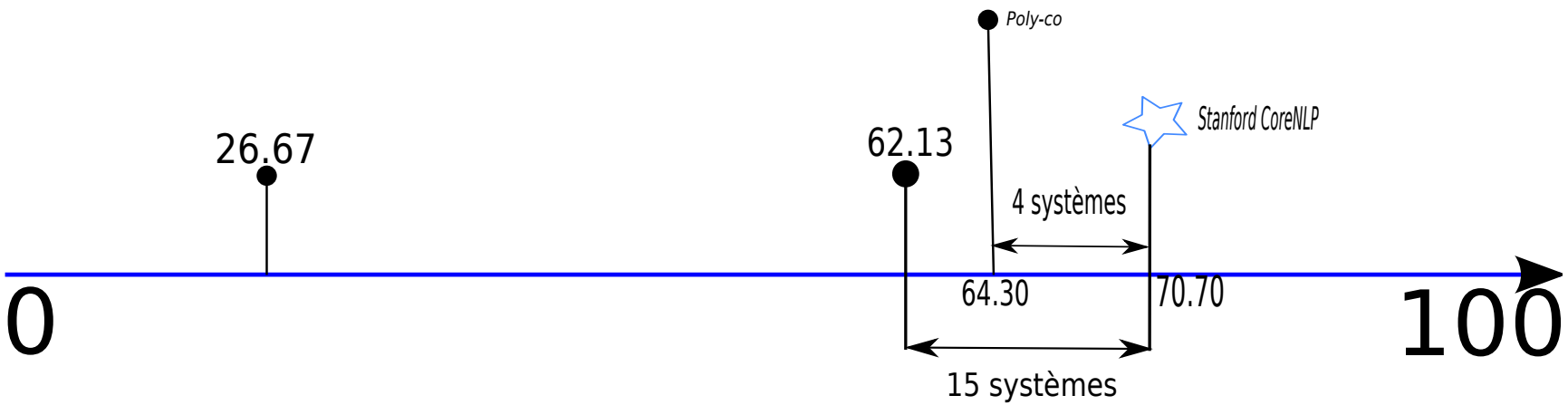
- Arbres: Nugues (C4.5), Uryupina,
- Perceptron : Stoyanov, Charton
- Maxent : Song, Santos, Kobdani
- SVM: Zhou

Résultats généraux



23 systèmes enregistrés, 18 systèmes publiés
11 pays - USA/5 - Chine/6 - Allemagne/3 - Brésil/2 -[Canada/1]

Détection des mentions



Détection des mentions - F-Score

Résultats officiels

System	MD	MUC	B-CUBED	$CEAF_m$	$CEAF_e$	BLANC	Official
	F	F^1	F^2	F	F^3	F	$\frac{F^1+F^2+F^3}{3}$
lee	**70.70**	59.57	68.31	**56.37**	**45.48**	73.02	**57.79**
sapena	43.20	59.55	67.09	53.51	41.32	71.10	55.99
chang	64.28	57.15	**68.79**	54.40	41.94	**73.71**	55.96
nugues	68.96	58.61	65.46	51.45	39.52	71.11	54.53
santos	65.45	56.65	65.66	49.54	37.91	69.46	53.41
song	67.26	**59.95**	63.23	46.29	35.96	61.47	53.05
stoyanov	67.78	58.43	61.44	46.08	35.28	60.28	51.92
sobha	64.23	50.48	64.00	49.48	41.23	63.28	51.90
kobdani	61.03	53.49	65.25	42.70	33.79	62.61	51.04
zhou	62.31	48.96	64.07	47.53	39.74	64.72	50.92
charton	64.30	52.45	62.10	46.22	36.54	64.20	50.36
yang	63.93	52.31	62.32	46.55	35.33	64.63	49.99
hao	64.30	54.47	61.01	45.07	32.67	65.35	49.38
xinxin	61.92	46.62	61.93	44.75	36.23	64.27	48.46
zhang	61.13	47.28	61.14	44.46	35.19	65.21	48.07
kummerfeld	62.72	42.70	60.29	45.35	38.32	59.91	47.10
zhekova	48.29	24.08	61.46	40.43	35.75	53.77	40.43
irwin	26.67	19.98	50.46	31.68	25.21	51.12	31.28

Système de référence

Stanford's Multi-Pass Sieve Coreference Resolution System
Heeyoung Lee, Yves Peirsman, Angel Chang, Nathanael Chambers, Mihai Surdeanu, Dan Jurafsky

Principes

- Système de *tamis* multiples inspiré de **Raghunathan et al. (2010)**.
- Repose intégralement sur des règles.

Algorithme

Algorithme

À chaque étape, un modèle de *tamis* gère un groupe de mentions. La précision de ce groupe est améliorée à chaque fois en donnant priorité aux mentions qu'il contient déjà.

1. Le premier tamis est celui de détection des mentions avec leurs caractéristiques (genre, nombre, etc).
2. Tamis de résolution des co-références.
3. Post-processing (ex: retrait des singletons).

Algorithme

Particularités des tamis de gestion de co-références

13 tamis successifs. Principaux tamis de résolution de co-référence:

1. (+) Par comparaison de similarité de chaînes nominales
2. (+) Par nom propre en tant que tête de mention
3. (-) Par incompatibilité des modifieurs (exemple : Liban $\not\sim$ Sud Liban)
4. (-) Par incompatibilité numérique (ex people $\not\sim$ few people)
5. (-) Si la distance entre un pronom et son antécédent $>$ n
6. (+) Similarité sémantique (comparaison des mots)
7. (+) Alias (comparaison des mots, avec utilisation de ressources)

Discussion

Comparaison avec les systèmes *classiques*

Le système de Stanford intègre une suite de modules heuristiques (les tamis) que l'on retrouve séparément dans tous les autres systèmes à classifieurs (architecture Soon).

- Question: quelle est l'influence réelle de la méthode de Soon ?

Discussion

Comparaison avec les systèmes *classiques*

Le système de Stanford intègre une suite de modules heuristiques (les tamis) que l'on retrouve séparément dans tous les autres systèmes à classifieurs (architecture Soon).

- Question: quelle est l'influence réelle de la méthode de Soon ?

Expérience

Retirer le classifieur de Poly-co, et ne conserver que les modules mentions et assemblage:

- Résultats ... surprenants.
- Baisse très modérée des performances (publication à venir).

Conclusions

Le classifieur de paires n'est pas le point central du système:

- Les modules de réunion par similarité ou incompatibilité de mentions (ex: PERS ne peut co-référer avec ORG) ont une influence très importante sur le score final.
- Le classifieur qui ne fait que modéliser des règles simples devient moins important si les systèmes d'étiquetage et de détection de mentions sont performants.
- La richesse des annotation de corpus est probablement plus importante que le système de résolution de paires.

Les tâches évoluées de NLP deviennent des assemblages complexes de sous-systèmes (EN, POS, Syntaxe, Analyse et Étiquetage Sémantique ...)

Perspectives

Proposition: Poly-co 3

Proposition de système hybride de détection de co-références:

- Intégrant la totalité des tamis implémentés par le groupe NLP de Stanford.
- Utilisant le résultat du classifieur de paires non plus comme central mais comme une information parmis d'autres des tamis.
- Introduire des modules de détection de similarité incorporant des notions sémantiques (ex étiquetage sémantique ajouté aux couches du corpus).

Architecture idéale ? A vérifier et comparer avec Ontonotes.

Ressources

Shared Task CoNLL 2011

- Site de la campagne:
  http://conll.bbn.com/
- Téléchargement de Poly-co 2:
  https://code.google.com/p/polyco-2/
- Corpus Ontonotes 2.0 (non public):
  http://www.ldc.upenn.edu/

Fin