La tâche *Gene Regulation Network in Bacteria* – BioNLP-ST'13

Robert Bossy, Philippe Bessières, Claire Nédellec

Séminaire NetBio

Paris – 12 septembre 2013

NetBIO

Extraction d'information –

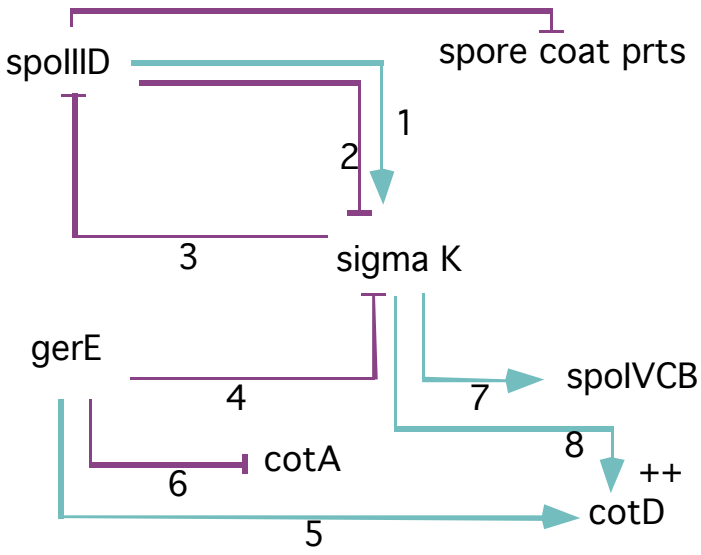
Construire des réseaux de régulation à partir de texte

positive interact.

negative interact.

1. SpoIIID is needed to produce sigma K

2. SpoIIID is capable of altering the specificity of RNAP-sigma K



3. Production of sigma K leads to a decrease in the level of spoIIID

4. GerE profoundly inhibits in vitro transcription of sigK encoding sigma K

5. GerE stimulates cotD transcription

6. ... and inhibits cotA transcription.

7. sigma K has been found that causes weak transcription of spoIVCB

8. ... and strong transcription of cotD

BioNLP Shared Task 2013

BioNLP, une série de compétitions internationales : 3 éditions en *2009*, *2011* and *2013*

Un cadre commun pour l'évaluation comparative de méthodes d'extraction d'événements précis à partir de textes, pour résoudre des problèmes du domaine biomédical

Principes

Extraction d'événements multiples et complexes à partir de documents scientifiques en langue naturelle

Focus sur la compréhension de la langue

De nombreuses tâches préparées par différents groupes et décrites dans un même langage formel.

Des questions biologiques variées en biologie moléculaire

- Des défis variés en Extraction d'Information (EI) contribuant à l'état de l'art.
- Pour encourager le développement de méthodes génériques

Organisation

De bonnes pratiques d'évaluation

Evaluation officielle sur 6 mois

Annotation de grande qualité en double-aveugle par des experts reconnus.

Calcul de l'accord inter-annotateur.

Etapes : Entraînement – Test-soumission – Rédaction d'article – ACL/HLT BioNLP workshop

Pour chaque tâche, les participants ont accès en ligne à

- La spécification de tous les événements avec leurs relations et les arguments valides
- Les données d'entraînement, de développement et de test.
- Les ressources linguistiques utiles au traitement (*Supporting resources*), principalement, les catégories morpho-syntaxiques et les dépendances (appel ouvert)
- Les programmes d'évaluation et d'analyse d'erreur.

Les services d'évaluation en ligne restent ouverts après la fin de l'évaluation officielle pour des comparaisons futures.

Série BioNLP-ST, résultats et participation

Résultats

Publications dans des numéros spéciaux de journaux d'informatique et de bioinformatique

Computational Intelligence (édition 2009) and *BMC Bioinformatics* (édition 2011)

De nombreux résultats sur les données BioNLP-ST publiés après la compétition.

Des progrès continus sont mesurés

Contribuant à la structuration de la communauté BioNLP

Participation

2009	24 équipes/participations à une tâche
2011	39 participations de 19 équipes aux 5 principales tâches (+13 participations à 3 *supporting tasks*)
2013	38 participations de 22 équipes aux 6 principales tâches

Edition 2013 de BioNLP Shared Task

- BioNLP-ST'13 suit les grandes lignes et objectifs des éditions précédentes.
- Le grand thème : *Knowledge base construction*
- Des questions biologiques et des espèces variées
- Différents types de documents, reflétant la diversité des sources de connaissances.
- De nouveaux défis en Extraction d'Information (IE) : Complexité croissante des événements à extraire. Etiqueter le texte avec une grande ontologie. Construction de réseau de relations.

6 tâches

- [GE] *Genia Event Extraction for NFkB knowledge base* (DBCLS)
- [CG] *Cancer Genetics* (NACTEM)
- [PC] *Pathway Curation* (NACTEM)
- [GRO] *Corpus Annotation with Gene Regulation Ontology* (NTU)
- [GRN] *Gene Regulation Network in Bacteria* (INRA)
- [BB] *Bacteria Biotopes* (INRA)

Trois tâches nouvelles (CG, PC and GRO) and 3 extensions de tâches précédentes (GE, GRN, BB)

[GRN] Gene Regulation Network in Bacteria

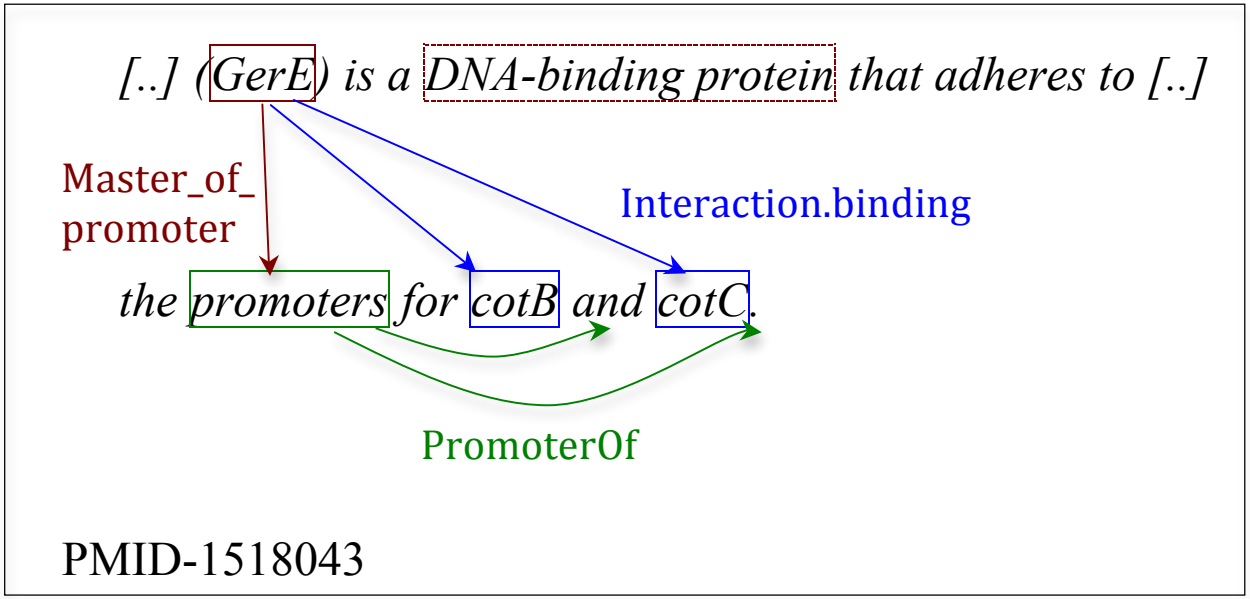
Spécificité des bactéries

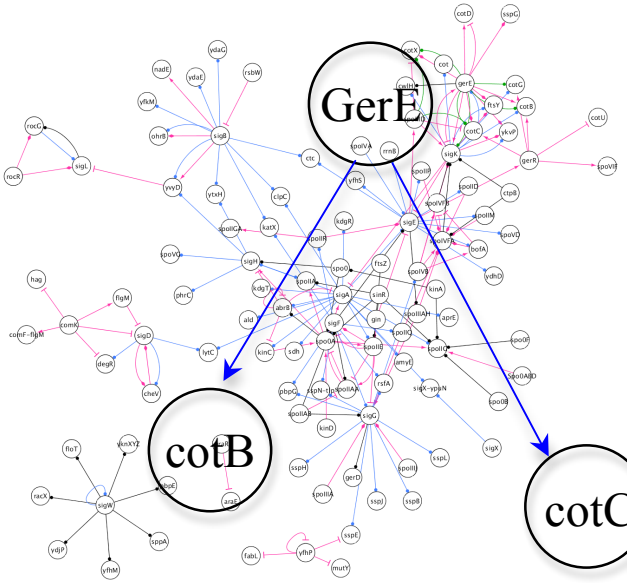
- Grande richesse des descriptions de régulations des bactéries modèles dans la littérature, principalement sur *Bacillus subtilis* et *Escherichia coli*.
- Peu de données d'interactions géniques structurées
- Des descriptions biologiques plutôt que physiologiques ou phénotypiques à la différence des eucaryotes

Objectif biologique

- Extraction du réseau de régulation d'un ensemble de gènes impliqués dans la sporulation de l'organisme modèle *Bacillus subtilis*.

Exemple





Les deux relations *Interaction* relient la protéine *GerE* aux gènes *cotB* et *cotC*

Historique de la tâche et données

Corpus annoté manuellement

Corpus de phrases extraites de références Pubmed répondant à la requête « Bacillus subtilis transcription sporulation ».

Annotations formelles par des bioinformaticiens et informaticiens : Philippe Bessières, Claire Nédellec, Erik Alphonse, Alain-Pierre Manine, Philippe Veber, Robert Bossy.

Annotations linguistiques par l'unité MIG.

Compétitions internationales

LLL *Learning Language in Logics* : une seule relation d'interaction entre protéines et gènes.

BioNLP ST'11 *Bacteria Interaction* : 13 entités et 10 relations.

BioNLP ST'13 *Genic Interaction Network in Bacteria*: 12 entités et 11 relations.

Modèle biologique, les entités

Entités biochimiques de la cellule bactérienne

Gene, *mRNA*, *Promoter*, *Protein* et *Site*.

Entités composites

GeneFamily: famille de gènes homologues.

Operon: opéron *sensu* procaryotes.

PolymeraseComplex: complexe RNA polymérase.

ProteinComplex: complexe protéique formé de protéines liées.

ProteinFamily: famille de protéines homologues.

Regulon: régulon, *sensu* procaryotes.

Modèle biologique, événements et relations biochimiques

Evénements

Transcription_by : transcription par la polymerase
Transcription_from : transcription à partir d'un site
Action Target : générique

Relations

Activation de promoteur

Promoter_of : relation entre gène et promoteur
Master_of_promoter : contrôle de la transcription depuis le promoteur par une protéine

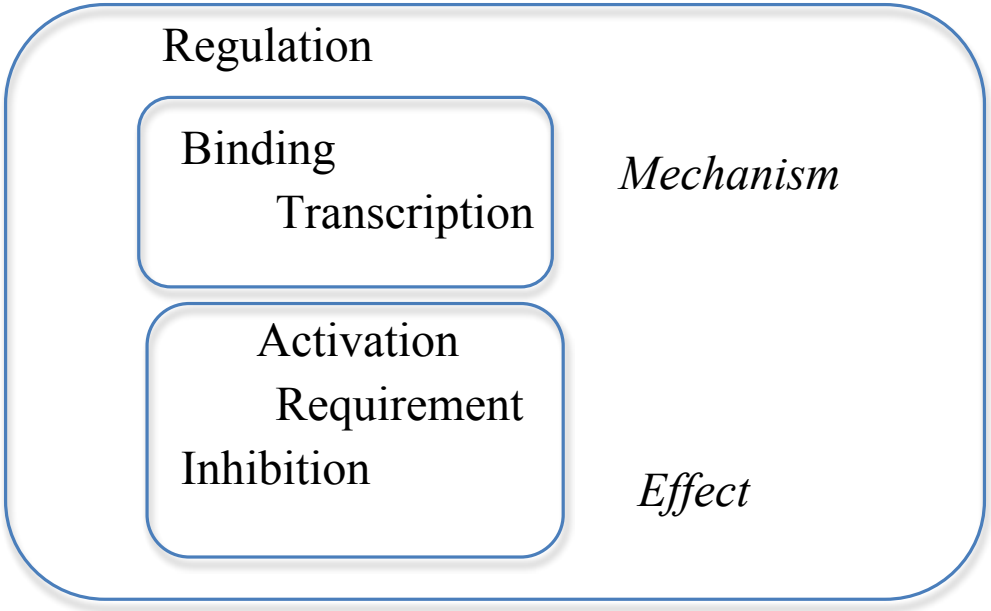
Régulons

Member_of_Regulon : appartenance du gène à un régulon
Master_of_Regulon : contrôle de l'activité du régulon par une protéine

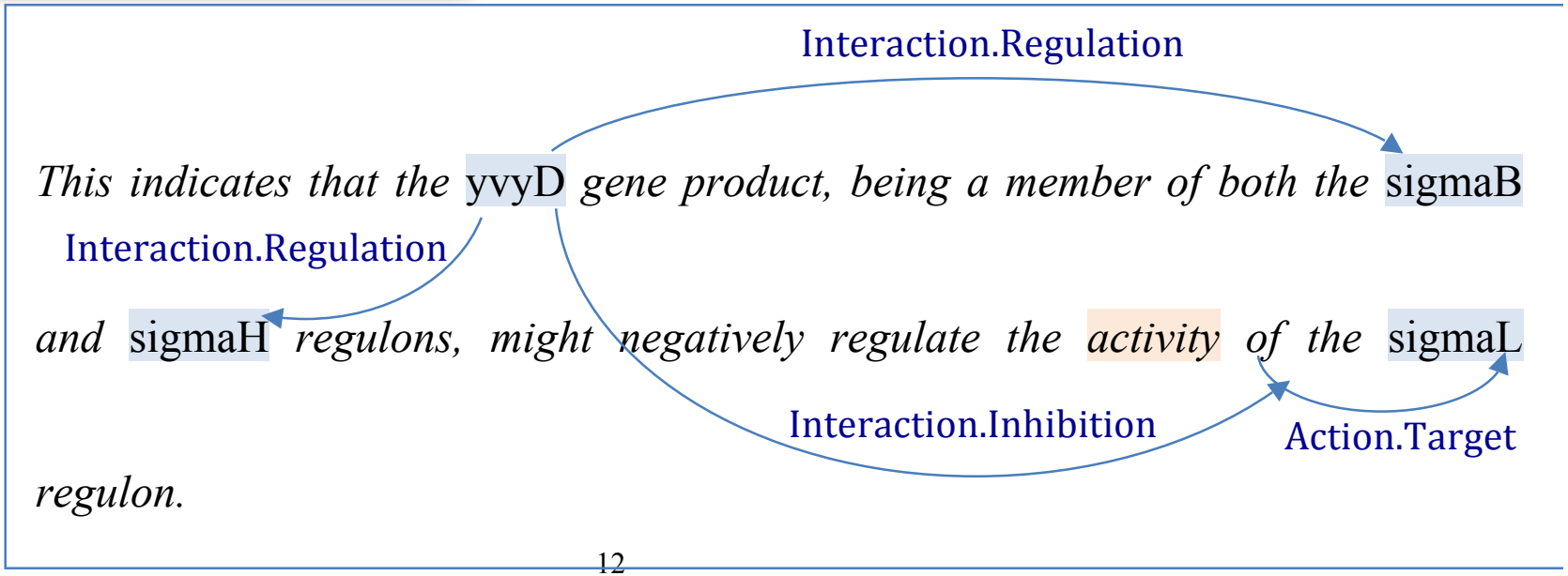
Liaison sur ADN

Bind_to : liaison de la protéine sur un site du chromosome
Site_of : appartenance d'un site à une entité génique (gène ou promoteur)

Modèle biologique, interactions



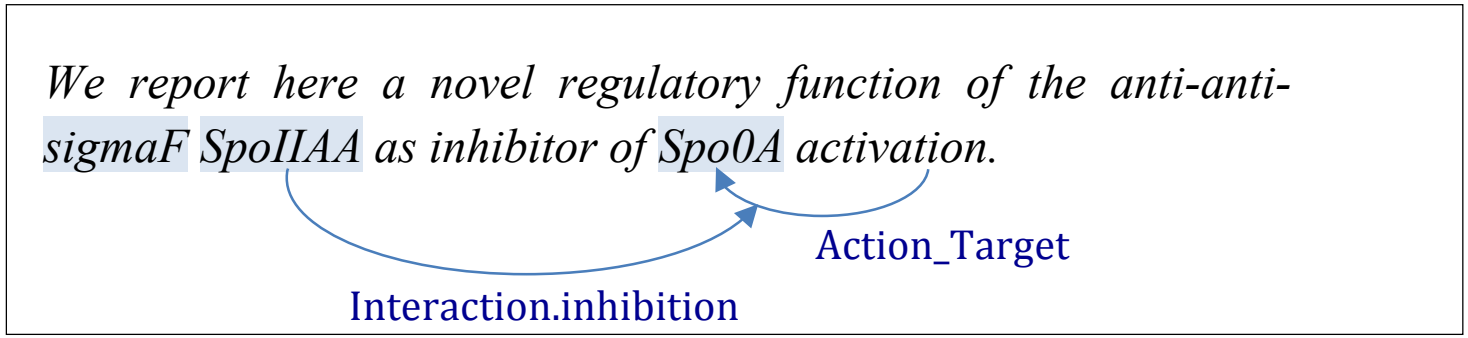
Exemple
PMID-9852014-txt



Modèle biologique, représentation

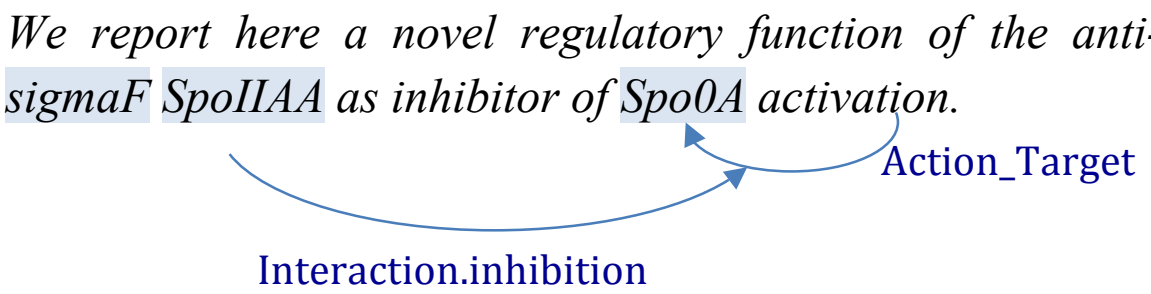
Les relations peuvent prendre comme arguments d'autres relations, ou des événements.

Exemple



Protein (SpoIIAA)
Protein (Spo0A)
Action Target (activation, Sp0A)
Interaction.Inhibition (SpoIIAA, Action Target (activation, Sp0A))

Langage formel, BioNLP



Les objets, entités, relations et événements sont identifiés par un identifiant unique, local à la phrase.
Les arguments des relations et des événements sont décrits par leur rôle et leur identifiant.

Les entités

Les entités sont décrites par leur type, leur position en caractères et leur texte.

T3 **Protein** 91 96 Spo0A

Les événements

Les événements sont décrits par leur type, le mot déclencheur et des arguments.

E1 **Action_Target:** T4 **Target:** T3

Les relations

Les relations sont décrites par leur type et leurs arguments.

R1 **Interaction.Inhibition Target:**E1 **Agent:** T2

Modèle biologique, le réseau

Le réseau biologique est déduit automatiquement des relations d'interactions extraites du texte et dans le format BioNLP

1. Sélection de relations d'interactions
2. Création des nœuds du réseau à partir des arguments des relations d'interactions
    a. Si argument = entités géniques ⇒ entités géniques
    b. Si argument = événement ⇒ entité génique de l'événement
    c. Si argument = relation ⇒ entités géniques de la relation.
    d. Si argument = entité promoteur ⇒ entité génique cible de la relation *Promoter_of*, ou entité génique agent de la relation *Master_of_Promoter*
3.
4. Suppression des arcs redondants. Le type le plus précis est conservé.

Exemple de dérivation du réseau de régulation génique à partir des faits extraits

We report here a novel regulatory function of the anti-anti-sigmaF SpoIIAA as inhibitor of Spo0A activation

Action_Target

Interaction.inhibition

R1 Interaction.Inhibition Target:E1 Agent:T2

T2 Protein 67 74 SpoIIAA

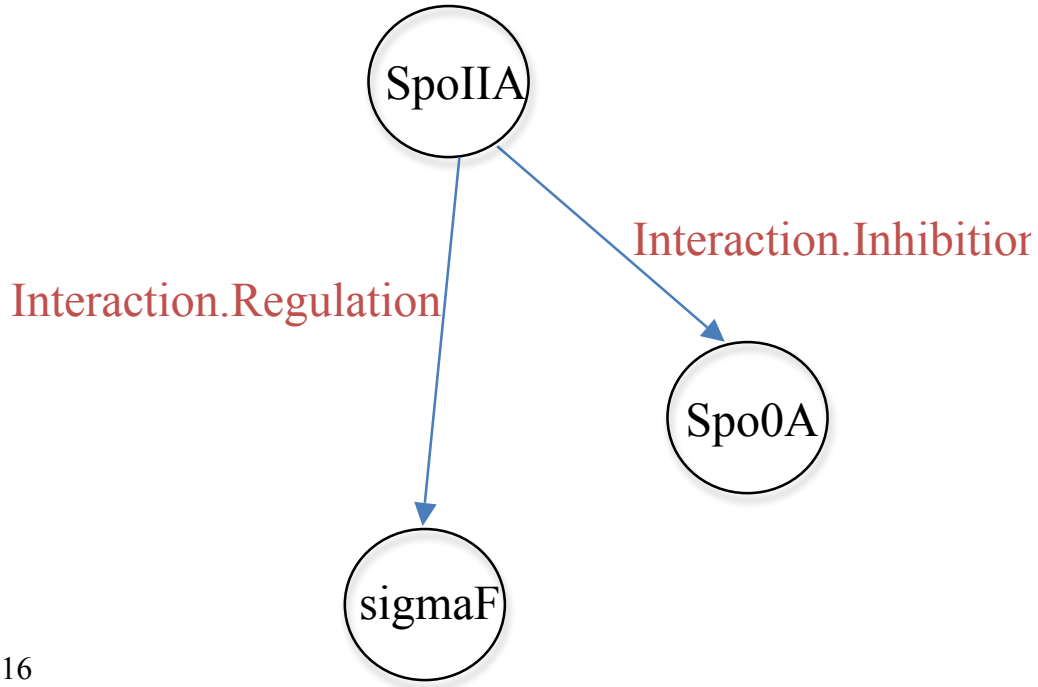
E1 Action_Target: T4 Target:T3

T4 Action 97 107 activation

T3 Protein 91 96 Spo0A

R2 Interaction.Regulation Target:T2 Agent:T1

T1 Protein 60 66 sigmaF



Les données de GRN

Les participants téléchargent

- Les textes des phrases
- Les annotations
- Les règles de dérivation du réseau

Phrases	201
Entités	1 161
Relations	499
Evénements	330
Nœuds	220
Arcs (159 de LLL)	440

Un tiers des phrases est utilisé pour l'évaluation (données de test)

Entités

Entité	#
Gene	199
GeneFamily	2
mRNA	1
Operon	33
PolymeraseComplex	62
Promoter	63
Protein	486
ProteinComplex	7
ProteinFamily	18
Regulon	14
Site	32
Total	**917**

Evénements biochimiques et relations

Evénements / Relations	#
Action target	226
Bind to	9
Master of Promoter	60
Master of Regulon	13
Member of Regulon	12
Promoter of	47
Site of	24
Transcription by	86
Transcription from	18
Total	**495**

Relations d'interaction

Interaction	#
Regulation	80
Inhibition	50
Activation	49
Requirement	35
Binding	12
Transcription	108
Total	**334**

Méthode d'évaluation

- Comparaison du réseau prédit et du réseau de référence.

- Utilisation de la mesure SER (*Slot Error Rate*) plus adaptée que la F-mesure pour éviter de pénaliser les substitutions deux fois.

$$SER = (S + D + I) / N$$

où

S est le nombre de substitutions (*i.e.* arcs prédits avec le mauvais type)
D est le nombre de suppressions est le nombre de suppresssions (faux négatifs)
I est le nombre d'insertions est le nombre d'insertions (faux positifs)
N est le nombre de d'arcs dans le réseau de référence

Résultats

Participant	SER	Rappel	Précision	SER *Shape*	SER *Effect*
U. of Ljubljana	0.73	34%	68%	0.60	0.74
K.U. Leuven	0.83	23%	50%	0.64	0.83
TEES-2.1	0.86	23%	54%	0.74	0.84
IRISA-TexMex	0.91	41%	40%	0.51	0.87
EVEX	0.92	13%	44%	0.79	0.91

La source d'erreur est principalement le type des arcs.

La relaxation des types surtout (SER *Shape*) ou des mécanismes (*Effect*) montre une nette amélioration.

Méthodes

Tous les participants calculent des informations linguistiques fournies :
lemmes, étiquettes morpho-syntaxiques et dépendances syntaxiques.

Toutes les méthodes utilisent un algorithme d'apprentissage.
Il est appliqué à un chemin syntaxique entre les arguments candidats, à l'exception de K.U. Leuven.

Participant	Méthode d'apprentissage
U. Ljubljana	Linear-chain CRF
K.U.Leuven	SVM (Gaussian RBF)
TEES-2.1	$SVM^{multiclass}$ (linear)
IRISA-TexMex	kNN (language model)
EVEX	SVM (TEES-2.1)

Toutes les tâches de BioNLP-ST'13

Tâches	Résultats d'évaluation (F-mesure, sauf quand SER)
GE 1 *Genia Event: Core event extraction*	TEES-2.1, EVEX, BioSEM: .51
GE 2 *Genia Event: Event enrichment*	TEES-2.1: .32
GE 3 *Genia Event: Negation/Speculation*	TEES-2.1, EVEX: .25
CG *Cancer Genetics*	TEES-2.1: .55
PC *Pathway curation New*	NaCTeM: .53
GRO *Gene Regulation Ontology*	TEES-2.1: .22 (*events*), .63 (*relations*)
GRN *Gene Regulation Network*	U. of Ljubljana: .73 (SER)
BB 1 *Bacteria Biotope*: *Entity detection and categorization*	IRISA: .46 (SER)
BB 2 *Bacteria Biotope*: *Relation extraction*	IRISA: .40
BB 3 *Bacteria Biotope*: *Full event extraction*	TEES-2.1: .14

Le système TEES-2.1 a participé à toutes les tâches sauf *BB 1*.
Il obtient les meilleurs résultats dans 6 tâches sur 9.

Discussion

- **Les méthodes sont de plus en plus généralistes**
  - Grâce à l'apprentissage automatique et à l'analyse des dépendances syntaxiques
  - Pas d'investissement dans le développement de ressources *ad'hoc*.
  - Des résultats encourageants sur des tâches difficiles

- Sur une **représentation « simple »** (relation entre protéines et gènes),

  LLL : 75 % à 80 % de F-mesure (rappel, précision)
  Des résultats utilisables automatiquement à grande échelle.

- Avec une **représentation très riche**, précision de l'ordre de 70%, mais rappel médiocre (34 %).
  Marge de progrès des méthodes
  - Meilleure utilisation des relations hiérarchiques
  - Augmenter le nombre de données d'apprentissage
  - Ne conserver de la représentation que les informations utiles aux biologistes

Conclusion

En mesurant la qualité des réseaux et non des informations extraites,
La tâche GRN ouvre la voie à une meilleure intégration des méthodes
de modélisation de réseaux et d'extraction d'information à partir de texte.

Services et données restent en accès public à des fins de comparaison dans la durée.

Données et description de la tâche : http://2013.bionlp-st.org/tasks/gene-regulation-network

Articles publics en ligne : http://www.aclweb.org/anthology-new/W/W13/ - 2000