Modélisation du dialogue:

Systèmes de dialogue parlé et corpus multilingues

Pierre Lison
Norwegian Computing Center (NR)
plison@nr.no

Séminaire du CENTAL

04/05/2018

Partie 1: Systèmes de dialogue parlé

Systèmes de dialogue

- = agents artificiels capables de communiquer avec des utilisateurs via les langues naturelles
- Langue parlée ou écrite (ex: chatbot), + autres modalités
- Souvent dans le cadre d'une tâche précise

Reconnaissance vocale

Analyse sémantique

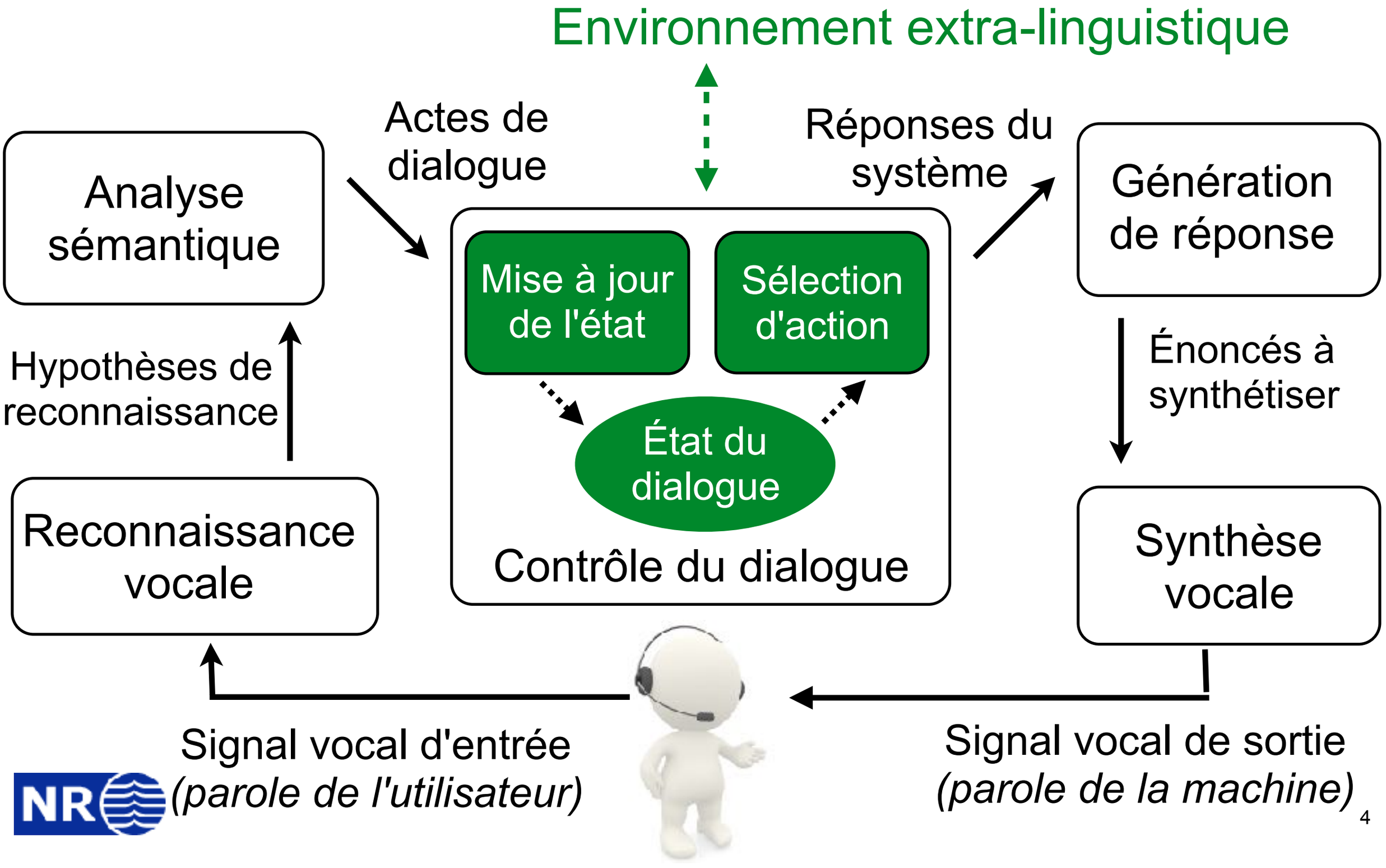
Contrôle du dialogue

Génération de réponse

Synthèse vocale

...

Architecture



Contrôle du dialogue

Approches symboliques

+ Maîtrise fine de l'interaction

− Prise en compte limitée des incertitudes

Approches statistiques ou neuronales

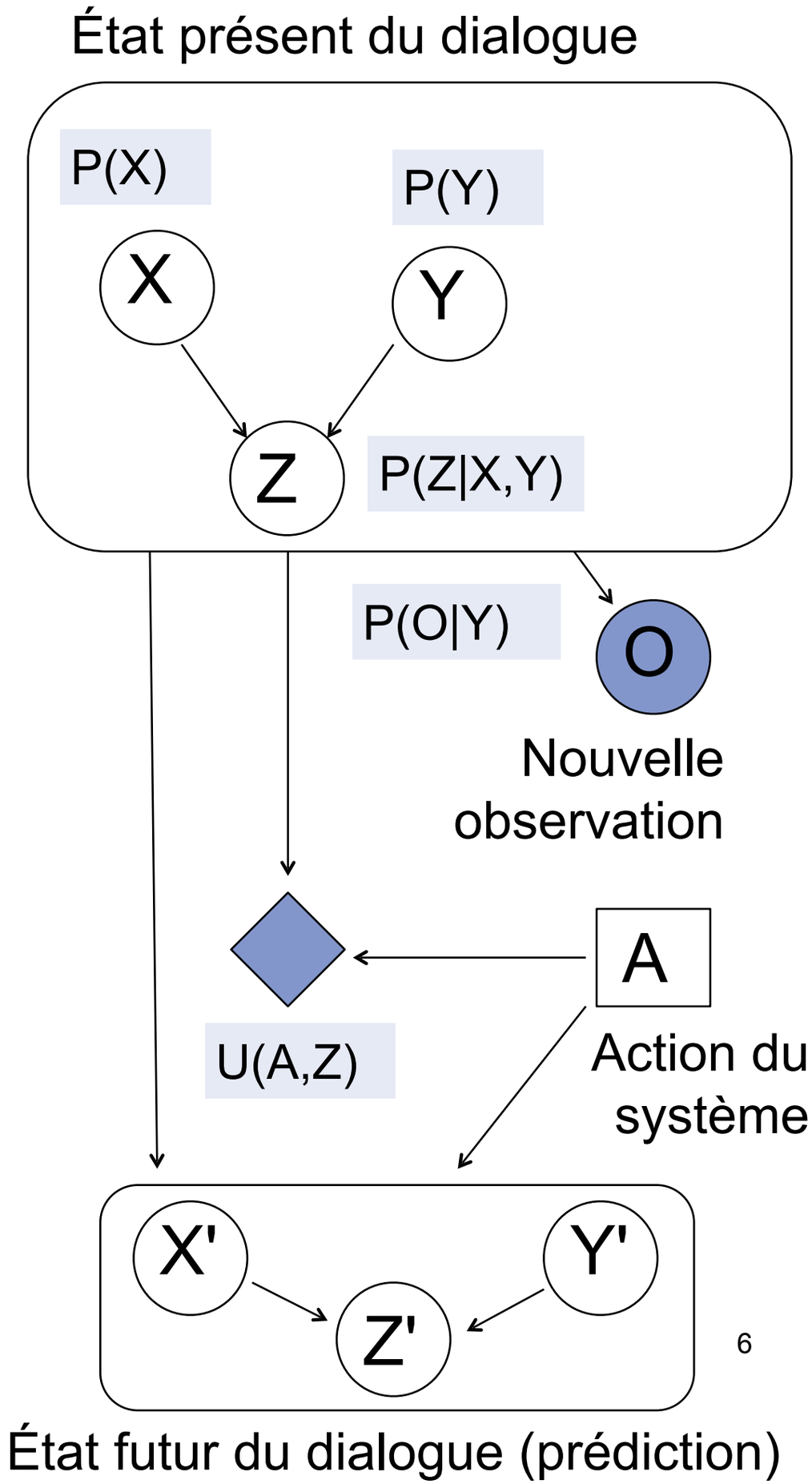
Modèles robustes, empiriques

Nécessitent de grandes quantités de données pour l'estimation

Approche hybride combinant modèles probabilistes, connaissance d'experts et petites quantités de données

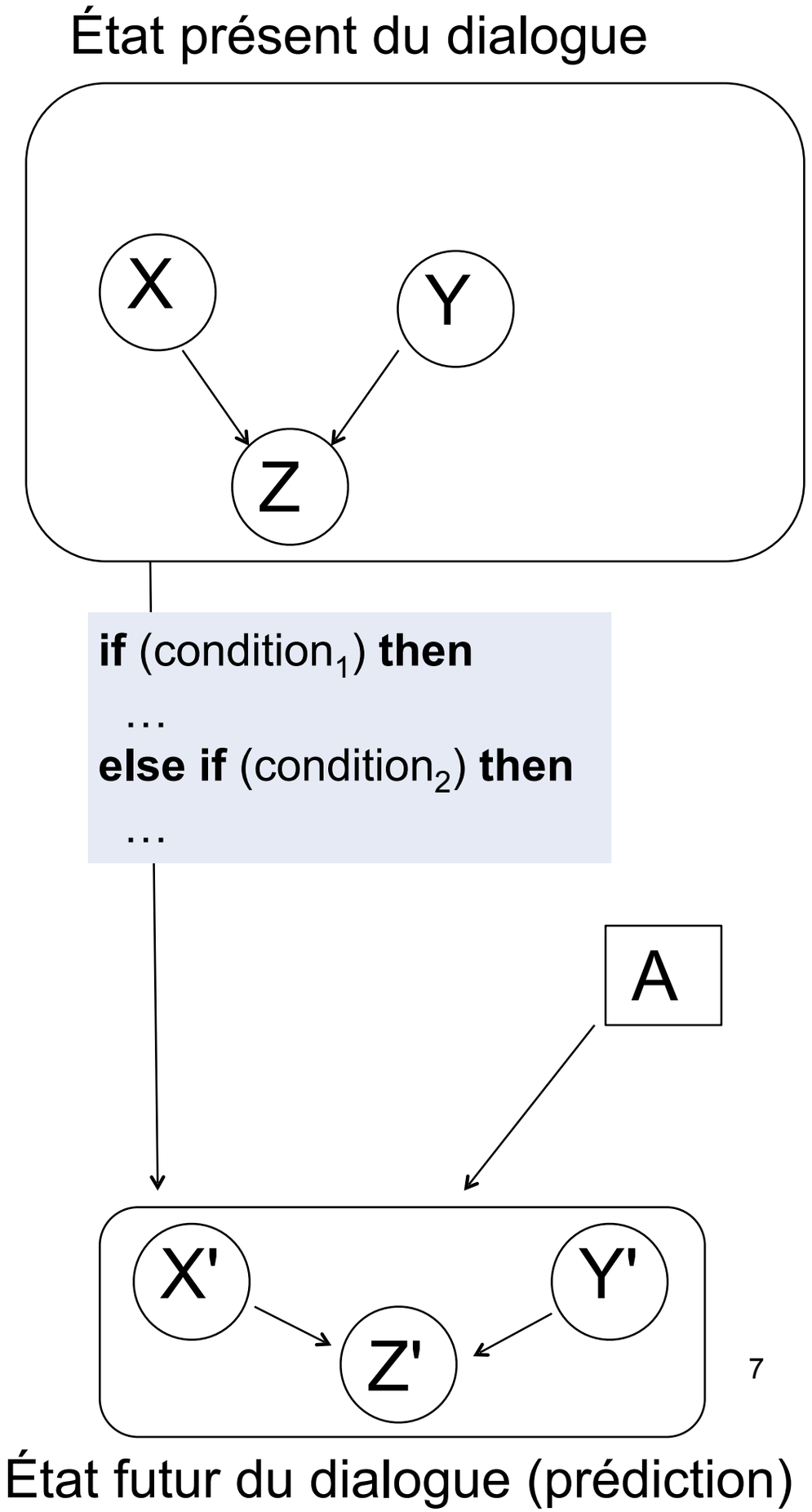
Idée en bref

- Représentation **probabiliste** de l'état du dialogue
  - Réseau bayésien où chaque variable exprime un aspect particulier de l'interaction
- Régulièrement mis à jour par de nouvelles observations (par ex. nouveaux actes de dialogue)
- Utilisé pour sélectionner des réponses du système
- (et prédire les états futurs)



Idée en bref

- **Problème**: grand nombre de paramètres à estimer
- **Solution**: représentation des modèles de dialogue via des *règles probabilistes*
- Introduit un *niveau d'abstraction* (logique) supplémentaire par dessus les modèles probabilistes classiques
- Permet de réduire grandement le nombre de paramètres



Structure des règles

if (condition$_1$ on **X**) **then**

P(**Y**=val$_1$) = θ_1
P(**Y**=val$_2$) = θ_2
…

else if (condition$_2$ on **X**) **then**

P(**Y**=val$_3$) = θ_3

P(**Y**=val$_4$) = θ_4

…

else

…

- Squelette en "*if-then-else*"
- *Conditions*: Formule logique sur des variables d'état X
- *Effets*: Attribution de valeurs sur des variables d'état Y
- Peut inclure des opérateurs logiques, des quantificateurs, etc.
- La structure des règles est fixée à l'avance, mais les paramètres θ peuvent être appris

Structure des règles

if (condition$_1$ on **X**) **then**

P(**Y**=val$_1$) = θ_1
P(**Y**=val$_2$) = θ_2
…

else if (condition$_2$ on **X**) **then**

P(**Y**=val$_3$) = θ_3

P(**Y**=val$_4$) = θ_4

…

else

…

- Lors de l'exécution, les règles sont *instanciées* comme variables dans un réseau Bayésien
- Permet d'utiliser les algorithmes d'inférence classiques
- Les règles probabilistes fonctionnent comme des "templates" de haut-niveau pour un modèle probabiliste traditionnel

Deux exemples

$\forall$ x,
if (*last-user-act* = x $\wedge$ *system-action* = AskRepeat) **then**
P(*next-user-act* = x) = 0.9

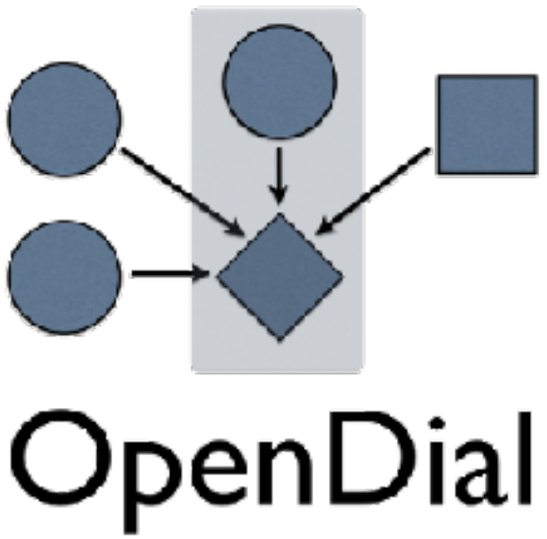
“Si le système demande à l'utilisateur de répéter son dernier acte de dialogue x, celui-ci devrait s'y conformer et répéter x avec une prob. 0.9.

$\forall$ x,
if (*last-user-act*=Request(x) $\wedge$ x $\in$ *perceived-objects*) **then**
U(*system-action*=PickUp(x)) = +5

"Si l'utilisateur demande au système de saisir un objet x et x est perçu par le système, l'utilité de saisir l'objet x est de 5. "

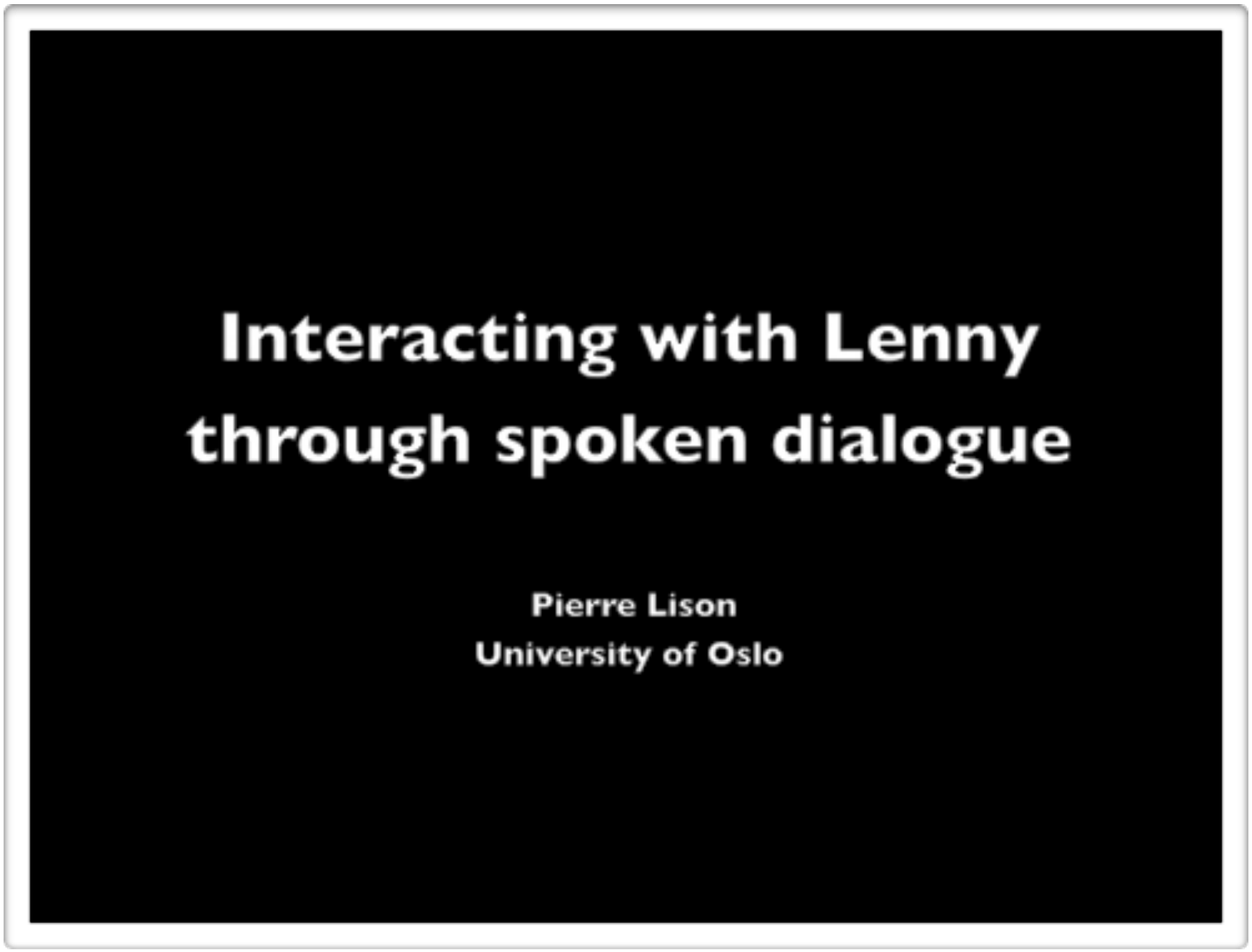
OpenDial

- Une boîte à outils pour le développement des systèmes de dialogue
  - Modèles du dialogue spécifiés par règles probabilistes
  - Plugins pour la reconnaissance vocale, l'analyse syntaxique, etc.
- Déployé dans plusieurs domaines:
  - Interactions homme-robot
  - Assistants de navigation multi-modaux
  - Guide culturel en ligne



Lison, P. and Kennington, C. (2016). *OpenDial: A toolkit for developing spoken dialogue systems with probabilistic rules*. ACL.

Interactions homme-robot



Pour plus de détails sur le protocole expérimental:

Lison, P. (2015). A hybrid approach to dialogue management based on probabilistic rules. *Computer Speech & Language*, 34(1):232-255

Partie 2: Corpus multilingues

Travail conjoint avec Jörg Tiedemann, Raveesh Meena & Seza Doğruöz

Sous-titres de films et de séries TV

Une ressource très intéressante pour le TALN:

1. Large spectre de genres linguistiques, expressions familières, conversations complexes, etc.
2. Grand volume de données disponibles (plusieurs millions de sous-titres téléchargeables)
3. Relation étroite entre les sous-titres et leur "source" → facilite l'alignement entre langues
4. Peut être associés avec des meta-data, des signaux audiovisuels, etc.

OpenSubtitles 2018

opensubtitles
.org

- Nous venons de publier une nouvelle version du corpus OpenSubtitles:
  - 3.35 milliards de phrases (22 milliards de mots) couvrant non moins de 60 langues
  - Les sous-titres sont alignés au niveau des documents et des phrases pour toutes les paires de langue
  - Actuellement la plus grande collection de corpus parallèles dans le domaine public

Disponible sur OPUS:
http://opus.nlpl.eu/OpenSubtitles2018.php

NR

Données initiales

Les administrateurs de www.opensubtitles.org nous ont aimablement fourni un "dump" de leur base de données

- 4 millions de sous-titres + méta-donnés (langue, format, identifiant IMDB, évaluations, etc.)
- Les sous-titres sont structurés en **blocs**, qui sont de petits segments de textes associés avec un temps de début et de fin (en millisecondes).
- Contraintes de temps et d'espace!

```
5
00:01:15,200 --> 00:01:20,764
Nehmt die Halme, schlagt sie oben ab,
entfernt die Blätter

6
00:01:21,120 --> 00:01:24,090
und werft alles auf einen Haufen
für den Pflanztrupp.

7
00:01:24,880 --> 00:01:30,489
Das Zuckerrohr beißt euch nicht.
Nicht so zaghaft! Na los, Burschen, los!
```

Quelques statistiques

Langue	Nombre de sous-titres	Nombre de phrases
Arabe	94.1K	83.6M
Bulgare	108K	94.6M
Croate	126K	113M
Tchèque	157K	136M
Néerlandais	125K	105M
Anglais	447K	441M
Français	127K	107M
Grec	143K	126M
Hébreu	98.7K	87.5M
Hongrois	131K	104M
Italien	135K	105M
Polonais	279K	237M
Portuguais	131K	118M
Portuguais (BR)	289K	252M
Roumain	205K	193M
Espagnol	234K	214M
Serbe	180K	168M
Turc	189K	173M

Prétraitement

1. Conversion vers Unicode

5
00:01:15,200 --> 00:01:20,764
Nehmt die Halme, schlagt sie oben ab,
entfernt die Blätter

6
00:01:21,120 --> 00:01:24,090
und werft alles auf einen Haufen
für den Pflanztrupp.

7
00:01:24,880 --> 00:01:30,489
Das Zuckerrohr beißt euch nicht.
Nicht so zaghaft! Na los, Burschen, los!

Prétraitement

1. Conversion vers Unicode
2. Segmentation en phrases

5
00:01:15,200 --> 00:01:20,764
Nehmt die Halme, schlagt sie oben ab,
entfernt die Blätter

6
00:01:21,120 --> 00:01:24,090
und werft alles auf einen Haufen
für den Pflanztrupp.

7
00:01:24,880 --> 00:01:30,489
Das Zuckerrohr beißt euch nicht.
Nicht so zaghaft! Na los, Burschen, los!

La détection des marqueurs de fin de phrases s'opère langue par langue (dépend du système d'écriture, des marques de ponctuation, etc.)

Prétraitement

1. Conversion vers Unicode
2. Segmentation en phrases
3. Tokenisation

5
00:01:15,200 --> 00:01:20,764
Nehmt die Halme , schlagt sie oben ab ,
entfernt die Blätter

6
00:01:21,120 --> 00:01:24,090
und werft alles auf einen Haufen
für den Pflanztrupp .

7
00:01:24,880 --> 00:01:30,489
Das Zuckerrohr beißt euch nicht .
Nicht so zaghaft ! Na los , Burschen , los !

- script tokenizer.perl de Moses
- Bibliothèques jieba and kytea pour le japonais et le chinois

Prétraitement

1. Conversion vers Unicode
2. Segmentation en phrases
3. Tokenisation
4. Correction des erreurs OCR

```
4
00:01:34,000 --> 00:01:38,471
<i>"Along with the sImpllflcatlon
I sought in my flrst films,</i>
5
00:01:39,080 --> 00:01:40,991
<i>"I wanted to be revolutlonary,</i>
```

- Approach "Noisy-channel" avec Google N-grams (11 langues européennes)
- 9 millions de mots corrigés

Prétraitement

1. Conversion vers Unicode
2. Segmentation en phrases
3. Tokenisation
4. Correction des erreurs OCR
5. Identification de la langue

Lui & Baldwin (2012) langid.py: An Off-the-shelf language identification tool, ACL 2012

Prétraitement

1. Conversion vers Unicode
2. Segmentation en phrases
3. Tokenisation
4. Correction des erreurs OCR
5. Identification de la langue
6. Extraction de méta-données

- **Infos sur la source:** Année de sortie, langue, durée, genre, ...
- **Infos sur le sous-titre:** Date d'upload, évaluations, durée, ..
- .
- **Infos sur la conversion:** Encodage, nombre de phrases, de mots, ...

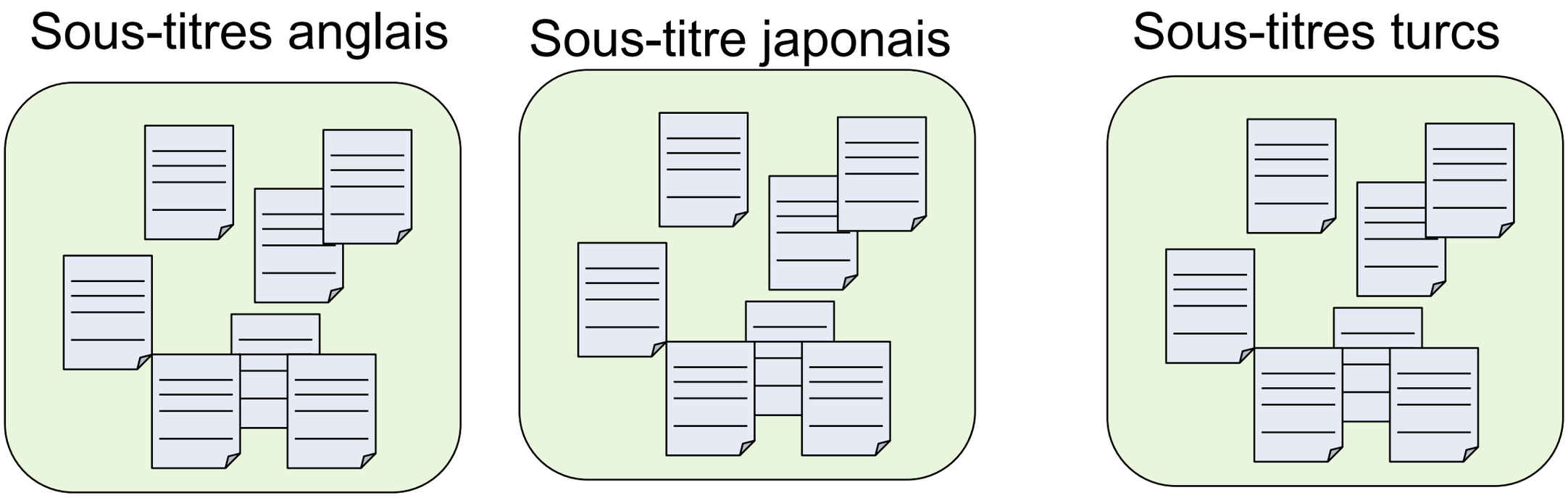
Prétraitement

1. Conversion vers Unicode
2. Segmentation en phrases
3. Tokenisation
4. Correction des erreurs OCR
5. Identification de la langue
6. Extraction de méta-données
7. Génération des fichiers XML

- **Infos sur la source:** Année de sortie, langue, durée, genre, ...
- **Infos sur le sous-titre:** Date d'upload, évaluations, durée, ..
- .
- **Infos sur la conversion:** Encodage, nombre de phrases, de mots, ...

Alignement

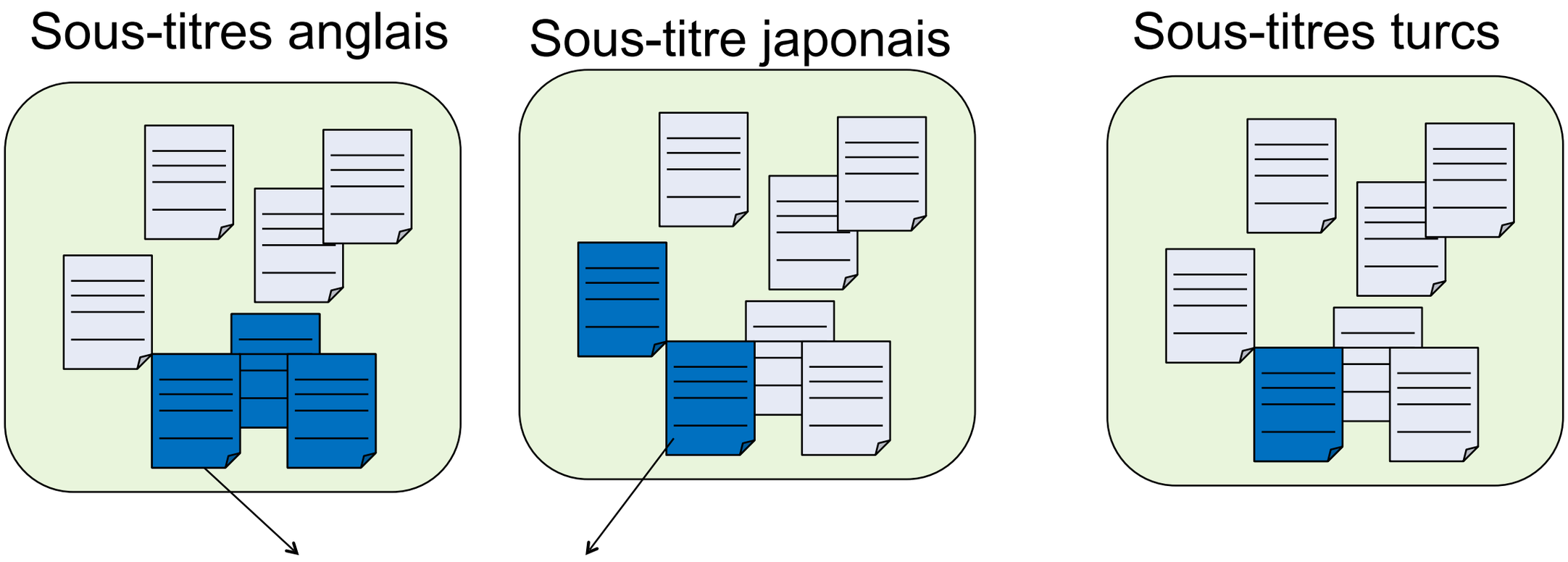
- Les sous-titres sont alors alignés entre paires de langues pour construire des corpus parallèles

- 

Alignement

= Sous-titres pour "Love actually" (2003), (basé sur l'identifiant IMDB)

- Les sous-titres sont alors alignés entre paires de langues pour construire des corpus parallèles

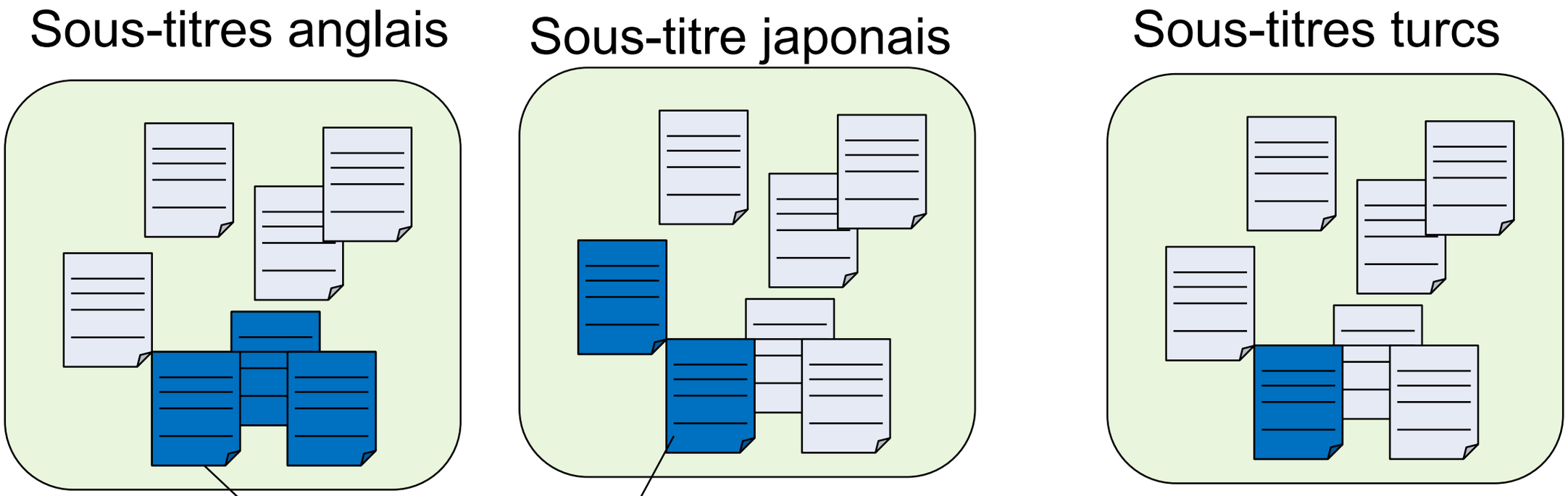
- 

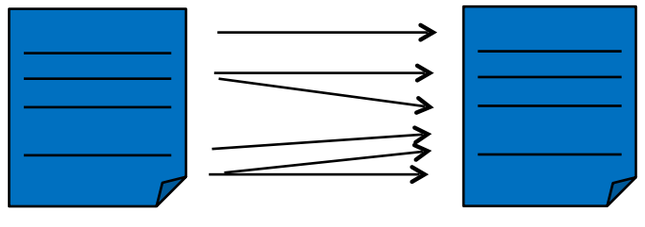
Fonction de score pour déterminer les "meilleures" paires de sous-titres (basé sur des mesures de qualités + recouvrement temporel)

Alignement

= Sous-titres pour "Love actually" (2003), (basé sur l'identifiant IMDB)

- Les sous-titres sont alors alignés entre paires de langues pour construire des corpus parallèles





Génération des **alignements de phrases** basées sur les temps d'affichage:

- Interpolation des temps inconnus
- Ratio de vitesse et décalage sont ajustés en utilisant des points d'ancrage

Extensions / améliorations

- Segmentations en tours de parole

  [Lison, P. & Meena, R. (2016), Automatic Turn Segmentation of Movie & TV Subtitles. *SLT* 2016.]

- Construction d'un modèle statistique permettant d'estimer la qualité de paires de phrases

  [P. Lison, J. Tiedemann & M. Kouylekov (2018), OpenSubtitles2018: Statistical Rescoring of Sentence Alignments in Large, Noisy Parallel Corpora, *LREC* 2018.]

- Détection des sous-titres de piètre qualité

  [P. Lison and A. S. Dogruöz (2018), Detecting Machine-translated Subtitles in Large Parallel Corpora, *BUCC* 2018.]

- Développement de modèles conversationnels neuronaux

  [Lison, P. & Bibauw, S. (2017), Not All Dialogues are Created Equal: Instance Weighting for Neural Conversational Models. SIGDIAL 2017.]

Segmentation en tours de parole

ID	Phrase	Début	Fin
1	If we wanted to kill you, Mr Holmes, we would have done it by now.	01:17:34.76	01:17:37.75
2	We just wanted to make you inquisitive.	01:17:37.80	01:17:40.59
3	Do you have it?	01:17:42.40	01:17:43.91
4	Do I have what?	01:17:43.91	01:17:45.43
5	The treasure.	01:17:45.48	01:17:46.43
6	I don't know what you're talking about.	01:17:46.43	01:17:48.91
7	I would prefer to make certain.	01:17:48.96	01:17:52.03
8	Everything in the West has its price.	01:17:57.00	01:17:59.63
9	And the price for her life - information.	01:17:59.68	01:18:04.55

Conclusions

- La modélisation du dialogue fait partie intégrante de nombreuses applications en TALN
- **OpenDial**: modèles de dialogue combinant apprentissage statistique et connaissances d'experts
  - Modèles + robuste (prise en compte des incertitudes), mais sans nécessiter de grands volumes de données
- **OpenSubtitles**: corpus multilingue construit à partir de sous-titres de films et séries TV
  - Les sous-titres sont alignés au niveau des documents et des phrases pour toutes les paires de langues