20/05/2021

# Analyzing complexity factors for Spoken Language Understanding on benchmark and deployed service corpora

**Rim Abrougui**

**TGI/DATA&AI/AITT/Deskiñ**

orange™

# INTRODUCTION

## Context & Problems

- Spoken Language Understanding models, involving contextual embeddings, have achieved remarkable results.
- Some SLU benchmark corpora remain challenging and performance can be affected by many factors related to the data (size, quality, annotation, ambiguity, etc.)

- ➢ How can we measure the complexity of corpora?
- ➢ What are the complexity factors that still resists to Transformers-based-models?
- ➢ Can this complexity be predictable when dealing with a new corpora ?
- ➢ Can data be partitionned into several sets representing different sources and levels of complexity?

# INTRODUCTION

## Objectives

- Measure the quality of a corpus and understand why it is difficult or easy
- Identify complexity factors that can be applied to any SLU task regardless of language, topic or semantic model linked to a given corpus.
- See how the DJINGO_SPK corpus is positioned in relation to public corpora used in the state of the art.


– Béchet, F., & Raymond, C. (2018). Is ATIS Too Shallow to Go Deeper for Benchmarking Spoken Language Understanding Models? *Interspeech 2018*, 3449-3453. https://doi.org/10.21437/Interspeech.2018-2256
– Béchet, F., & Raymond, C. (2019). Benchmarking Benchmarks : Introducing New Automatic Indicators for Benchmarking Spoken Language Understanding Corpora. *Interspeech 2019*, 4145-4149. https://doi.org/10.21437/Interspeech.2019-3033
– Bechet, F., Raymond, C., Hamane, A., Abrougui, R., Marzinotto, G., & Damnati, G. (2021). *Analyzing complexity factors for Spoken Language Understanding on benchmark and deployed service corpora*. 5.


**Submission Interspeech 2021**

# APPROACHES: First Step

1. Select a set of corpora with concept's annotation in word level
2. Train independently a set of DNN models on the datasets for concept prediction
3. Labeling each token in the test part of corpora with labels according to **agreements** and **correctness**:

- ❖ if all the models agree on the same prediction, then the tokens will have the label "**Agreement**" (A), otherwise, they belong to "**No Agreement**" (N)
- ❖ if at least one algorithmic system predicts the correct label, then the tokens will have the label "**Correct**" (C), otherwise they belong to "**Error**" (E)

# APPROACHES: First Step

## 4 Clusters

- ❖ AC: Agreement + Correct
- ❖NC: No agreement + Correct
- ❖AE: Agreement + Error
- ❖NE: No agreement + Error

| $i$ | word $w_i$ | label(ref,u,i) | label($m_1$,u,i) | label($m_2$,u,i) | cluster |
|---|---|---|---|---|---|
| 1 | find | O | O | O | **AC** |
| 2 | flights | O | O | O | **AC** |
| 3 | arriving | O | O | O | **AC** |
| 4 | new-york | B-to-city | B-to-city | **B-from-city** | **NC** |
| 5 | new-york | O | **B-to-city** | **B-to-city** | **AE** |
| 6 | next | B-date-arr | **B-date-dep** | **O** | **NE** |
| 7 | saturday | I-date-arr | **I-date-dep** | **B-date-arr** | **NE** |

Béchet, F., & Raymond, C. (2019). Benchmarking Benchmarks : Introducing New Automatic Indicators for Benchmarking Spoken Language Understanding Corpora. *Interspeech 2019*, 4145-4149. https://doi.org/10.21437/Interspeech.2019 3033

## 2 levels of difficulty

- ❖ AC: Agreement + correct => easy samples (all tokens of a sample have the label AC)
- ❖ NCE: tous les autres exemples = > difficult samples (at least one token of a sample has the label NCE)

# DATASETS

| | Domain | Language | Collect Method | Annotation |
|---|---|---|---|---|
| ATIS | Information on flights, airlines, airports | English | Wizard approach & manual transcription of recordings. Some automatic approaches were realized on new version of ATIS | Annotation with semantic frames (Intent & slots) accroding to relational schema BIO encoding |

| Token | Label de référence |
|---|---|
| i'm | O |
| traveling | O |
| to | O |
| Dallas | B-toloc.city_name |
| from | O |
| Philadelphia | B-fromloc.city_name |

1) Tur, G., Hakkani-Tur, D., & Heck, L. (2010). What is left to be understood in ATIS? 2010 IEEE Spoken Language Technology Workshop, 19-24. https://doi.org/10.1109/SLT.2010.5700816
2) Dahl, D. A., Bates, M., Brown, M., Fisher, W., Hunicke-Smith, K., Pallett, D., Pao, C., Rudnicky, A., & Shriberg, E. (1994). Expanding the scope of the ATIS task : The ATIS-3 corpus. Proceedings of the Workshop on Human Language Technology - HLT '94, 43. https://doi.org/10.3115/1075812.1075823

# DATASETS

| | Domain | Language | Collect Method | Annotation |
|---|---|---|---|---|
| SNIPS | Weather information, restaurant booking, Music, etc.<br>7 tasks: SearchCreativeWork, GetWeather, BookRestaurant, PlayMusic, AddToPlaylist, RateBook, SearchScreeningEvent | English, French, German, Spanish and Korean. | ASR system + manual verification | Manual annotation by Amazon Mechanical Truck crowdsourcing<br>Annotation in intents & concepts<br>BIO encoding |

| Token | Label de référence |
|---|---|
| add | O |
| another | O |
| song | O |
| to | O |
| the | O |
| Cita | B-playlist |
| Romantica | I-playlist |
| playlist | O |

Coucke, A., Saade, A., Ball, A., Bluche, T., Caulier, A., Leroy, D., Doumouro, C., Gisselbrecht, T., Caltagirone, F., Lavril, T., Primet, M., & Dureau, J. (2018). Snips Voice Platform : An embedded Spoken Language Understanding system for private-by-design voice interfaces. ArXiv:1805.10190 [Cs]. http://arxiv.org/abs/1805.10190

# DATASETS

| Corpus | Domaine | Langue | Méthode de collecte | Annotation |
|---|---|---|---|---|
| M2M | A fusion of two datasets containing dialogues for restaurant and movie ticket booking. | English | Automatic dialogue & crowdsourcing<br>(1) Providing a task schema and an API client<br>(2) Generation of dialogue outlines<br>(3) Rewriting the utterances and validating slot spans<br>(4) Training a dialog model with supervised learning on the dataset. | • Automatic reading of the dialogues generated between 2 chatbots (BU & BS) generating a sequence of annotations for each round of dialogue.<br>• Dialogue frame annotation encoding the dialogue act sentence (intent) and a slot value<br>• Repeat the process until the user's goals are met and the user exits the dialog with a "bye ()" act, or a maximum number of turns is reached.<br>• The remaining rounds of dialogue are annotated with two simpler crowdsourcing tasks: "Does this utterance contain this particular location value?" and "Do these two statements have the same meaning?"<br>BIO encoding |

| Token | Label de référence |
|---|---|
| the | O |
| date | O |
| is | O |
| this | O |
| wednesday | date-B |
| at | O |
| the | O |
| camera | theatre_name-B |
| 7 | theatre_name-I |



Shah, P., Hakkani-Tür, D., Tür, G., Rastogi, A., Bapna, A., Nayak, N., & Heck, L. (2018). Building a Conversational Agent Overnight with Dialogue Self-Play. ArXiv:1801.04871 [Cs]. http://arxiv.org/abs/1801.04871

# DATASETS

| | Domaine | Langue | Méthode de collecte | Annotation |
|---|---|---|---|---|
| MEDIA | Touristic information | French | Collection by ELDA Wizard of Oz approach The recording platform included an automatic generator to help agents with their responses. | • Manual transcription and annotation with concepts according to a rich semantic ontology + BIO encoding<br>• The semantic dictionary used for the annotation of the MEDIA corpus associates with a word or a group of words a concept-value pair then a specifier defining relations between concepts and finally an affirmative, negative, interrogative or possible mode, attached to the concept.<br>Word \| concept c \| mode \| spécifieur \| valeur<br>**Je voudrais reserver** \| commande \| +(affirmatif) \|  - \| reservation<br>**une chambre** \| chambre-quantite \| + \| reservation \| 1<br>**pour deux nuits** \| sejour-nbNuit \| + \| reservation \| 2 |

| Token | Label |
|---|---|
| ça | O |
| fait | O |
| à | B-comparatif-paiement |
| à | I-comparatif-paiement |
| peu | I-comparatif-paiement |
| près | I-comparatif-paiement |
| combien | B-objet |
| pour | O |
| une | B-nombre |
| chambre | B-chambre-type |
| simple | B-chambre-type |



Devillers, L., Maynard, H., Rosset, S., Paroubek, P., McTait, K., Mostefa, D., Choukri, K., Charnay, L., Bousquet, C., Vigouroux, N., Béchet, F., Romary, L., Antoine, J. Y., Villaneau, J., Vergnes, M., & Goulian, J. (2003). The French MEDIA/EVALDA project : The evaluation of the understanding capability of Spoken Language Dialogue Systems. 4.

# DATASETS

| Corpus | Domaine | Langue | Méthode de collecte | Annotation |
|---|---|---|---|---|
| DJINGO_SPK | Set of skills and interactions with corporate services (Orange TV, music with its partner Deezer, Orange Radio, telephony), general services (weather forecast, shopping, calendar, news) and general interaction with customers (small interviews, global orders). | French | Real logs ASR transcription | Semantic annotations directly performed on ASR transcriptions |

| Token | Label de référence |
|---|---|
| [CLS] | Music_play |
| je | O |
| veux | O |
| une | O |
| playlist | B- designator_playlist |
| de | O |
| rock | B-playlist |
| road | I-playlist |
| trip | I-playlist |
| [CLS] | Smart_Home_Turn_On |
| la | B-object_name |
| lumière | I-object_name |

# DATASETS' CHARACTERISTICS

| Corpus _test | Djngo_SPK | Atis | Media | Snips |
|---|---|---|---|---|
| #word | 34938 | 8333 | 25977 | 6595 |
| #sent | 9984 | 893 | 3005 | 700 (100 sent par intent) |
| Vocabulary | 2637 | 485 | 1219 | 1752 |
| #concept | 34 | 84 | 70 | 39 |
| #intent | 109 | - | - | 7 |
| %OOD sentences | 6.6% | 0 | 0 | 0 |
| %sent in train | 76.9% | 1.9% | 44.6% | 0.9% |
| %sent with concept | 59.3% | 99.3% | 86.5% | 100% |
| Av sent length | 4.2 | 10.3 | 7.6 | 9.16 |
| Concept av_length | 1.5 | 1 | 1.99 | 1.77 |

**Caractéristiques du corpus DJINGO**

- The biggest
- Most frequent number of intents and least frequent number of concept
- Out-of-domain sentences
- The majority of sentences are seen in train data
- Unequal distribution of concepts over sentences
- The shortest sentences
- Compound concepts (B + I) less frequent than Media and Snips and more frequent than Atis.

# EXPERIMENTS

## Models

DJINGO_SPK

| pretraining | self attention | bigru | lstm | fine-tuning |
| --- | --- | --- | --- | --- |
| DistilBERT | M1 | M2 | M3 | |
| CamemBERT | | | | M4 |

BENCHMARK CORPORA

| pretraining | bigru | gru | self attention |
| --- | --- | --- | --- |
| BERT | M1 | M3 | M5 |
| random | M2 | M4 | M6 |

Table 2: *Description of models M1 to M6 in terms of pretraining conditions and DNN architecture*

# APPROACHES: Second Step

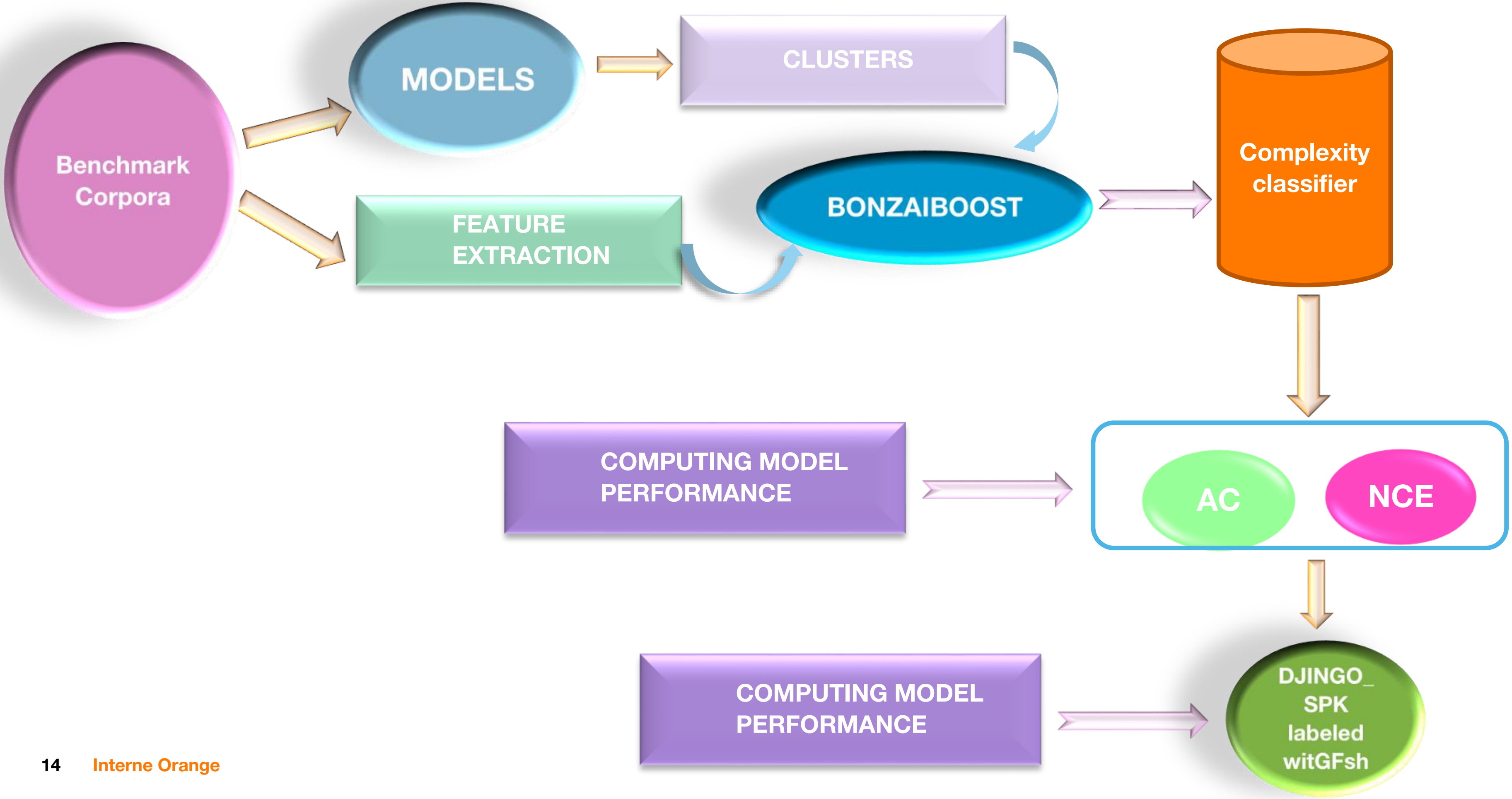
1. Describe each word in the test corpora of each SLU corpus with characteristics independent of language and subject and independent of the concept (Generic features GF)

2. Train a classifier on the corpora described by GFs to predict complexity labels (AC or NCE) (bonzaiboost: decision tree + boosting)

3. Evaluate the performance of a model on corpora distributed in AC and NCE (the labels predicted by the classifier)

4. Analyzing complexity factors in NCE

# APPROACHES



Benchmark Corpora
MODELS
CLUSTERS
FEATURE EXTRACTION
BONZAIBOOST
Complexity classifier
COMPUTING MODEL PERFORMANCE
AC
NCE
COMPUTING MODEL PERFORMANCE
DJINGO_ SPK labeled witGFsh

# COMPLEXITY FEATURES

**GF complexity categories:**

- **Ambiguity**: long statement, multiple verbs, disfluencies ...
- **Coverage**: OOV, rare association between token-label, new word n-gram

| |
|---|
| **Ambiguity** |
| # of semantic labels acceptable for $W$ |
| # of Part-Of-Speech (POS) acceptable for $W$ + POS label |
| # of possible syntactic dependency for $W$ + dependency label |
| distance between $W$ and the sentence syntactic root. |
| utterance length (in words) |
| $\%$ of words in $S$ belonging to a concept |
| **Coverage** |
| # of occurrences of $W$ in train |
| # of occurrences of $(W, l)$ in train |
| is bigrams $(W-1, W)$ and $(W, W+1)$ occurring in train? |

Table 1: *The Generic Feature (GF) set*

# BENCHMARK CORPORA RESULTS

## Models' performance

| development corpus | ATIS | MEDIA | SNIPS | M2M |
|---|---|---|---|---|
| #word | 8333 | 25977 | 6595 | 28119 |
| #sent | 893 | 3005 | 700 | 4800 |
| #concepts | 84 | 70 | 39 | 12 |
| *Concept detection performance for models M1... M6* | | | | |
| Fmes(M1,all) | 94.6 | 85.7 | 95.4 | 91.5 |
| Fmes(M2,all) | 93.8 | 81.7 | 69.6 | 91.7 |
| Fmes(M3,all) | 94.7 | 85.8 | 95.2 | 93.6 |
| Fmes(M4,all) | 79.0 | 60.1 | 69.0 | 91.0 |
| Fmes(M5,all) | 94.8 | 85.3 | 95.9 | 93.0 |
| Fmes(M6,all) | 77.4 | 59.8 | 68.9 | 91.0 |
| *Repartition into easy (AC) and difficult (NCE) sentences* | | | | |
| AC | 46.2% | 54.3 | 35.1 | 84.2 |
| NCE | 53.8% | 45.7 | 64.9 | 15.8 |
| *Performance of model M1 on AC and NCE sentences* | | | | |
| Fmes(M1,AC) | 98.7 | 98.5 | 99.7 | 99.0 |
| Fmes(M1,NCE) | 91.7 | 82.3 | 93.1 | 68.6 |

Table 3: *Corpora characteristics and concept detection performance for SLU models M1...M6 on all sentences (all), and model M1 on easy sentences (AC) and difficult sentences (NCE)*

=> Performance obtained with a state-of-the-art model (M1) is much worse on NCE utterances compared to AC utterances

# BENCHMARK CORPORA RESULTS

## Bonzaiboost classification performance

⇒ F-measure over 93% for label AC

⇒ F-measure almost 60% for label NCE

=> Encouraging results: Complexity labels were predicted with any lexical or semantic information

| ATIS | Precision | Recall | F-measure |
|---|---|---|---|
| *AC* | 91.75 | 98.26 | 94.89 |
| *NCE* | 60.61 | 23.26 | 33.61 |
| **MEDIA** | **Precision** | **Recall** | **F-measure** |
| *AC* | 82.55 | 87.82 | 85.11 |
| *NCE* | 63.03 | 52.80 | 57.46 |
| **SNIPS** | **Precision** | **Recall** | **F-measure** |
| *AC* | 92.54 | 96.04 | 94.26 |
| *NCE* | 58.93 | 42.31 | 49.25 |
| **M2M** | **Precision** | **Recall** | **F-measure** |
| *AC* | 98.08 | 99.89 | 98.98 |
| *NCE* | 97.00 | 65.10 | 77.91 |
| *All corpora* | | | |
| **all** | **Precision** | **Recall** | **F-measure** |
| *AC* | **91.58** | **95.57** | **93.53** |
| *NCE* | **68.42** | **52.21** | **59.23** |
| *All* | **88.83** | **88.83** | **88.83** |

Table 4: *Classification performance on AC/NCE labels with the GF feature set. Training on the union of all corpora.*

# BENCHMARK CORPORA RESULTS

## Analysis of NCE decisions in terms of the respective weights of the ambiguity and coverage features

⇒ Depending on the corpus considered, the complexity can come because of:

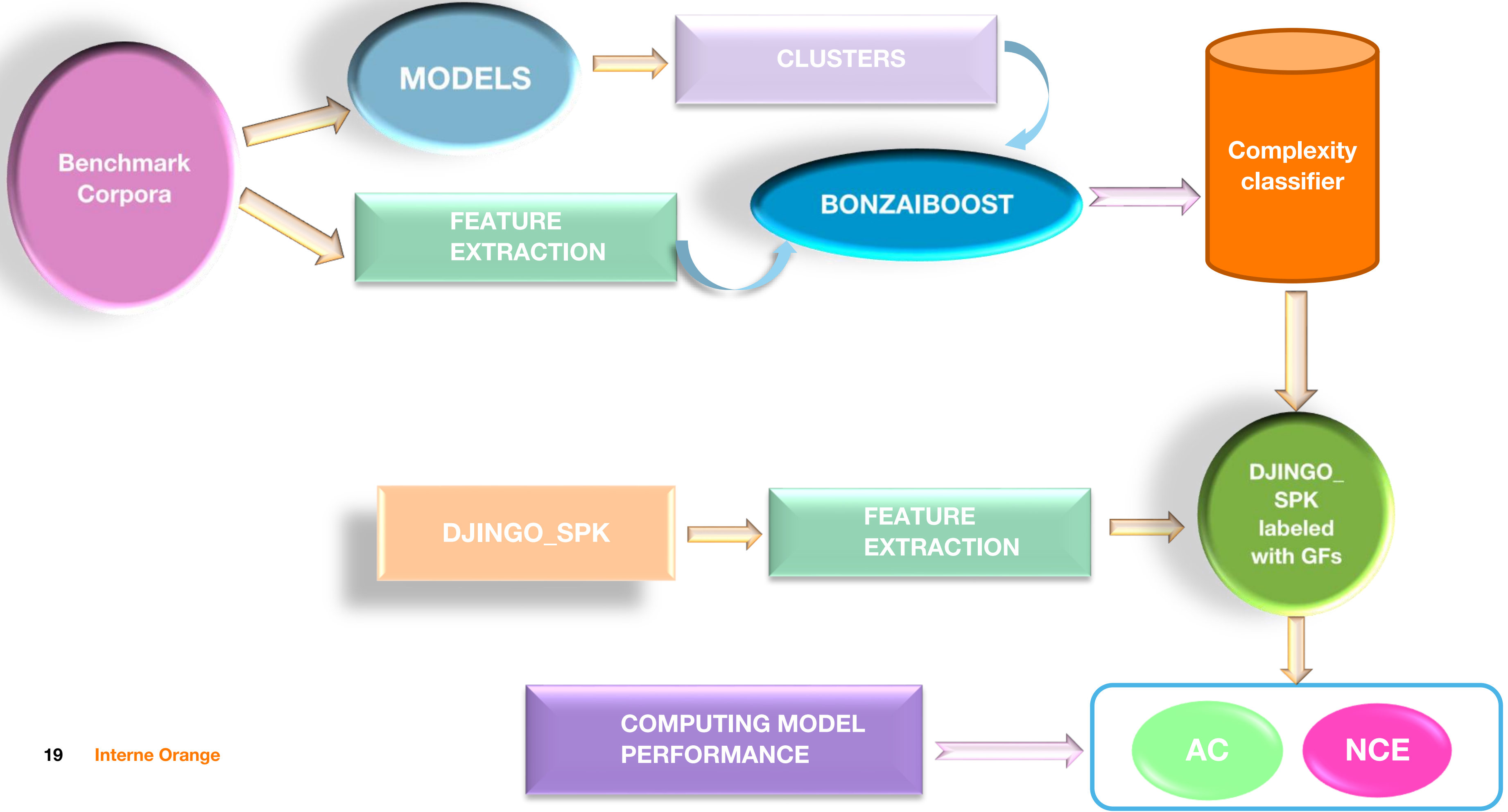
- ❑ Coverage issues (ATIS & M2M)
- ❑ Ambiguity issues (MEDIA)
- ❑ Coverage & Ambiguity (SNIPS)

=> The classifier can still be used to accurately partition a corpus according to criteria linked to the utterance complexity and the sources of this complexity.

| | | |
|---|---|---|
| **ATIS** | **weight(NCE,AMBIG)** | **weight(NCE,COVER)** |
| *reference* | 13.1% | 86.9% |
| *prediction* | 19.9% | 80.1% |
| **MEDIA** | **weight(NCE,AMBIG)** | **weight(NCE,COVER)** |
| *reference* | 84.4% | 15.6% |
| *prediction* | 84.3% | 15.7% |
| **SNIPS** | **weight(NCE,AMBIG)** | **weight(NCE,COVER)** |
| *reference* | 37.2% | 62.8% |
| *prediction* | 23.5% | 76.5% |
| **M2M** | **weight(NCE,AMBIG)** | **weight(NCE,COVER)** |
| *reference* | 4.1% | 95.9% |
| *prediction* | 2.3% | 97.7% |
| **all** | **weight(NCE,AMBIG)** | **weight(NCE,COVER)** |
| *reference* | **65.8%** | **34.2%** |
| *prediction* | **68.0%** | **32.0%** |

Table 5: *% of weight for boosting rules belonging to the* ***ambiguity*** *(AMBIG) category vs. the* ***coverage*** *(COVER) category.*

# APPLICATION ON DJINGO_SPK



# EVALUATION METHODS

## 3 possible levels to evaluate

1. Token's level

- **label correct = O or label of concept with borders (B and I) correct**
- label correct =  O or label of concept without borders correct
- **concatenated label correct = intent-O + intent-concept with borders correct**
- concatenated label correct =  intent-O + intent-concept without borders correct

2. Entity's level

- entity correct = concept correct
- entity correct = concept + intent correct
- **entity correct = concept + borders correct**
- **entity correct = (concept +borders correcte) + intent correct**

3. Sample's level

- **sample correct = intent correct**
- **sample correct = all the concepts + all the borders of a sample correct**
- **sample correct = (all the concepts + all the borders of a sample correct) + intent correct**

# EVALUATION METRICS

| Token's level |
| --- |
| **Accuracy**<br>**(nb labels corrects/nb tokens)** |

| Entity's level |
| --- |
| **Précision for each concept (P)**<br>**(nb concepts ok) / (nb concepts hyp)** |
| **Recall for each concept (R)**<br>**(nb concepts ok) / (nb concepts ref)** |
| **F1 mesure for each concept =**<br>$\frac{2*P*R}{R+P}$ |
| **F1 Macro** $=\frac{1}{N}\sum_{i=0}^{N} F1\ score$ |
| **F1 Micro =**<br>$\frac{2*P\_global*R\_global}{R\_global+P\_global}$ |

| Sample's level |
| --- |
| **Accuracy**<br>**(nb samples corrects/nb samples)** |

# COMPLEXITY FACTORS IN DJINGO_SPK

| Partition | AC | NCE | ALL |
|---|---|---|---|
| coverage | 86.5% | 13.5% | 100% |
| token accuracy | 98.6 | 92.4 | 97.3 |
| F1 concepts | 95.6 | 83.8 | 92.2 |
| sample accuracy (intents + concepts OK) | 95.7 | **79.7** | 93.5 |
| Weight (AMBIG) | - | 28.9 | - |
| Weight(COVER) | - | **71.1** | - |

- **16% drop between results on AC partition vs NCE**
- **Most of the complexity factors come from coverage issues**
- **Almost of 30% of the complexity factors come from ambiguity issues**

# COMPLEXITY FACTORS IN SLU CORPORA

| NCE Weight % | DJINGO_SPK | ATIS | SNIPS | MEDIA | M2M |
|---|---|---|---|---|---|
| Ambig | 28 | 19.9 | 23.3 | 84.3 | 2.3 |
| Cover | 71.1 | 80.1 | 76.5 | 15.7 | 97.7 |

**% weight of difficult utterances (NCE) - AMBIG vs COVER**

- **MEDIA > SNIPS > DJINGO_SPK > ATIS > M2M**

# Conclusion

- The Djingo corpus is different from other public SLU corpus since it is the result of a collection in real situation
- It is possible to measure the quality of the corpora and understand the complexity factors without retraining.
- We could analyze other complexity factors by adding other families to the GFs and conclude rules measuring the degrees of difficulty of NLU corpora.

# Merci

# DJINGO_SPK RESULTS

## Models' performance

| Token's evaluation level | Self-attention | Bi-Gru | LSTM | CamemBert |
|---|---|---|---|---|
| **Accuracy**<br>**label = O & concepts**<br>**(nb labels corrects/nb tokens)** | 96.94 | 96.85 | 96.70 | **97.31** |
| **Accuracy**<br>**Label = (O-intents) + (concepts-intents)**<br>**(nb labels correct/nb tokens)** | 92.7 | 92.31 | 92.58 | **93.19** |
| | | | | |

- ➢ Close results
- ➢ CamemBert has the best performance

# DJINGO_SPK RESULTS

## Models' performance

| Sample's evaluation level | Self-attention | Bi-Gru | LSTM | CamemBert |
|---|---|---|---|---|
| **Accuracy**<br>**Sample correct = intent correct**<br>**(nb samples corrects/nb samples)** | 96.33 | 95.87 | 96.36 | **96.48** |
| **Accuracy**<br>**Sample correct = all concepts + boundaries correct**<br>**(nb samples corrects/nb samples)** | 94.93 | 94.29 | 94.40 | **95.39** |
| **Accuracy**<br>**Sample correct = intent + all concepts + boundaries correct**<br>**(nb samples corrects/nb samples)** | 93.17 | 92.11 | 92.61 | **93.51** |

➢ Accuracy results at **intent+concepts+borders** level are the least efficient

# DJINGO_SPK RESULTS

## Models' performance

| Entity's evaluation level | Self-attention | Bi-Gru | LSTM | CamemBert |
|---|---|---|---|---|
| **F1 Macro (concept + frontière)** | 79.38 | 76.67 | 74.61 | **81.11** |
| **F1 Macro (concept + frontière + intent)** | 76.47 | 73.51 | 71.9 | **78.26** |
| **F1 Micro (concept + frontière)** | 91.93 | 90.34 | 90.42 | **92.22** |
| **F1 Micro (concept + frontière+intent)** | 88.57 | 86.61 | 87.14 | **88.98** |
| | | | | |



➢ L'évaluation au niveau entité est plus stricte que les deux autres évaluations

# Résultats d'un modèle entrainé sur les corpus SLU

- **réseau : self-attention**

| Corpus | ATIS | MEDIA | SNIPS | DJINGO |
|---|---|---|---|---|
| **Niveau token**<br>**Accuracy**<br>**label = O et concepts**<br>**(nb labels corrects/nb tokens)** | **97.7** | 89.6 | **97.8** | **96.9** |
| **Niveau sample**<br>**Accuracy**<br>**Sample correct = tous les concepts + frontières corrects**<br>**(nb samples corrects/nb samples)** | 88.1 | 76.1 | 90.3 | **91.9** |
| **Evaluation niveau concept**<br>**F1 Micro** | **94.8** | 85.3 | **95.9** | **94.9** |

- Résultats proches
- Les résultats du modèle entrainé sur le corpus MEDIA sont les moins bons

- Les résultats d'un même modèle entrainé sur chaque corpus et les métriques d'évaluation n'expliquent pas pourquoi un corpus est plus complexe ou plus difficile qu'un autre