

Document understanding as a way to improve search relevance

Pauline Chavallard, Nicolas Fiorini
26/11/19

Search Solutions 2019, London, UK



Plan



About Doctrine

The search challenge

Problem statement

Methods

Results

Conclusions & future work



Doctrine was created in 2016

Challenges

- volume of data - > search
- heterogeneity
- domain specificity



Product overview

BNP paribas

100 € offerts

Dossiers

Alertes

Nicolas

Décisions500 +

Commentaires406

Lois et règlements260

Avocats0

Entreprises500 +

PAR DATE

☒ Date indifférente

☐ Depuis 2019

☐ Depuis 2018

☐ Depuis 2015

☐ Période spécifique

☐ Date précise

RECHERCHE DANS LE TEXTE

☐ Exclure les moyens annexés

PAR JURIDICTION

▶ ☐ TGI

500 +

▶ ☐ Cour d'appel

500 +

▶ ☐ Tribunal de commerce

500 +

Offre de prêt

Opération de change

Taux effectif global

Contrat de crédit

Mise en état

Intérêts conventionnels

> M

Plus de 500 décisions (0,40 s)

Créer une alerte

Trier : Par pertinence ▼

Souhaitez-vous restreindre la recherche à l'entreprise : [BNP PARIBAS](#) ?

1

Cour de cassation, Chambre sociale, 6 juin 2018, 17-14.372 17-14.373 17-14.374 17-14.375, Publié au bulletin, société [BNP Paribas](#) c/ pôle 6

Décision attaquée : Cour d'appel de Paris, 8 décembre 2016 | Dispositif : Rejet

Contrat de travail, exécution · Intéressement · Participation · Salarié · Détachement · Succursale · Entreprise · Étranger · Dispositif · Accord collectif

... cassation ; PAR CES MOTIFS : REJETTE les pourvois ; Condamne la société [BNP Paribas](#) aux dépens ; Vu l'article 700 du code de procédure civile, la condamne ... été conclus pour les années considérées au sein du groupe [BNP Paribas](#) et que la [BNP Paribas](#) SA est l'un des signataires de ces accords ; que l'article ... » ; qu'il en résulte qu'à aucun moment, contrairement aux dires de [BNP PARIBAS](#), Monsieur Y n'a été, individuellement et de manière claire et expresse, ... dispositif de participation, il ressort des accords conclus au sein du groupe [BNP](#) les 30 juin 2000, 29 juin 2005, 30 juin 2009 et 30 juin 2010 (pièces n° 32, ... [Lire la suite...](#)

Salariés Expatriés [BNP Paribas](#) dans une succursale à l'étranger : droit à la participation et l'intéressement ! (c. cass. 6 juin 2018, n°17-14372)

Frederic Chhum · blogavocat · 23 juin 2018

Participation et intéressement : les salariés Expatriés à Londres, Singapour ou New York de [BNP Paribas](#) dans une succursale à l'étranger y ont droit !

Me Frédéric Chhum · consultation.avocat.fr · 23 juin 2018

[11 autres commentaires](#)

VOS FILTRES

Vous pouvez filtrer sur la chambre d'une juridiction pertinente pour vous, sur une entreprise, un avocat...

BNP PARIBAS

Appellation commerciale : **HELLO BANK!**

SIREN : **662 042 449**

Ville : **PARIS 9**

Secteur d'activité : **Autres intermédiations monétaires**

Taille : **10 000 salariés et plus**

Suivre cette entreprise

Filter sur les décisions

☐

JURIDICTIONS

VILLES

ÉVOLUTION



Legalese

“ And whereas we have consented to the said period being so brought to an end and to the exercise of such right of presentation now we hereby declare that the said period shall come to an end on the date hereof and that the said vacancy in the said Benefice of Ansdell and Fairhaven Saint Paul in Our said Diocese of Blackburn may thereupon be filled. ”

Source: Golden Bull Award

Plan

About Doctrine

The search challenge

Problem statement

Methods

Results

Conclusions & future work





Better search quality

Can be achieved with

- improved **ranking**
 - algorithm selection
 - feature design & dataset creation
- improved **query understanding**
 - user intent classification & question answering
 - automatic filters
- improved **document understanding**
 - segmentation
 - language models and summarization



Better search quality

Can be achieved with

- improved **ranking**
 - algorithm selection
 - feature design & dataset creation
- improved **query understanding**
 - user intent classification & question answering
 - automatic filters
- improved **document understanding**
 - segmentation
 - language models and summarization



Time consuming,
requires gold
standard



Better search quality

Can be achieved with

- improved **ranking**
 - algorithm selection
 - feature design & dataset creation
- improved **query understanding**
 - user intent classification & question answering
 - automatic filters
- improved **document understanding**
 - segmentation
 - language models and summarization

Time consuming,
requires gold
standard

Difficult without a
large user base or
expertise



Better search quality

Can be achieved with

- improved **ranking**
 - algorithm selection
 - feature design & dataset creation
- improved **query understanding**
 - user intent classification & question answering
 - automatic filters
- improved **document understanding**
 - segmentation
 - language models and summarization

Time consuming,
requires gold
standard

Difficult without a
large user base or
expertise

Best opportunity



Unlocks many solutions

- text matching — synonyms
- relevance ranking — segmentation
- salient fact identification (axioms, conclusions...)
- reference detection (e.g. to civil law articles)
- search filter on text part

Plan

About Doctrine

The search challenge

Problem statement

Methods

Results

Conclusions & future work



A french court decision is generally structured following these sections:

- **Metadata** (« En- tête » in French): court, number, date, etc., of the trial.
- **Parties** (« Parties » in French): information about the claimants and defendants
- **Composition of the court** (« Composition de la cour » in French)
- **Facts** (« Faits » in French): what happened?
- **Pleas in law and main arguments** (« Moyens » in French): arguments presented by the claimant and defendant.
- **Grounds** (« Motifs » in French): reasons and arguments used by the court
- **Operative part of the judgment** (« Dispositif » in French): final decision

Unfortunately, there is no mandatory guideline on how to release a court decision.

Courts may use:

- different styles in term of writing
- different styles in term of organising the documents
- all sections from previous slide, or a subset



The French Court of Appeal usually has a very unified way of writing: ~55 % have explicit titles for their categories

Facts

Exposé du **litige** :

Le 3 septembre 2003, Madame A B, épouse X, se trouvant au marché de XXX et C D, à Y, a buté sur le pied du parasol abritant le stand de **fruits** et légumes de Monsieur G E F et s'est blessée à l'épaule gauche en tombant au sol.

Extracted from <https://www.doctrine.fr/d/CA/Orleans/2007/SKDD824CCFE8D8D9D93128>.



First approach: pattern recognition

```
)
r"D\s*[ÉE]\s*C\s*I\s*D\s*E|O\s*R\s*D\s*O\s*N\s*N\s*E", TAGS_MAPPING["DISPOSITIF"])), # CE / CAA / TA:
r"(?i)(?:^(Les )?Motifs?(Les )?Motifs? de l['']arr[eê]t|(Les )?Motifs? de la (d[ée]cision|cour)"
r"|DISCUSSION.{0,20}|SUR CE ?,? ?(LA COUR)?|SUR QUOI ?,? ?(LA COUR)?|Expos[ée] d(es|u) motifs?.{0,5}$", TAGS_MAPPING["MOTIFS"])),
r"(?:^M O T I F S).{0,50}$", TAGS_MAPPING["MOTIFS"])),
r"(?i)(?:^Expos[ée].{0,5}$|Faits.{0,5}proc[ée]dure.{0,30}|Faits et pr[ée]tentions des parties"
r"|Expos[ée] (du litige|de l'affaire)|Rappel (des faits|de la proc[ée]dure).{0,30}|FAITS PROC[EÉ]DURE.{0,20}).{0,5}$", TAGS_MAPPING["FAITS"])),
r"(?:^F A I T S).{0,50}$", TAGS_MAPPING["FAITS"])),
r"(?i)(?:^Composition (de la cour|du tribunal).{0,50})$", TAGS_MAPPING["COMPOSITION"])),
r"(?i)(?:^D[EÉ]BATS ?.{0,5})", "Débats"),
r"(?i)(ENTRE.{0,5}|APPELANTE?S?.{0,5}|DEMANDE(UR|RESSE).{0,50}|Vu la proc[ée]dure entre|. {0,10}PARTIES.{0,30}"
r"|.{0,100}affaire entre.{0,5}|.{0,100}proc[ée]dure introduite par.{0,5})$", TAGS_MAPPING["PARTIES"])),
AS_TITLE, r"(?i).{0,5}\s*Sur l", None),
AS_TITLE, r"D[EÉ]CISION|FAITS|D[EÉ]BATS|COUR|SUR CE|PROC[EÉ]DURE|PR[ÉE]TENTIONS|MOYENS|"
r"PARTIES|DISPOSITIF|ARGUMENTS|DISCUSSION|EXPOS[ÉE] DU LITIGE$", None),
AS_TITLE, r"(?i)(PREMIER|DEUXIEME|SECOND|. {3,5}I[EÈ]ME) MOYEN DE CASSATION|CECI ([ÉE]TANT )?EXPOS[EÉ]", None),
H, NORMALIZE_TITLE, r"(?i)^Par ces motifs|^P A R C E S M O T I F S", TAGS_MAPPING["DISPOSITIF"])),
H, MATCH_ONLY_AS_TITLE, r"MOYENS? ANNEXES?", None),
H, ALL_PARAGRAPH_AS_TITLE, r"(?i)^(Mais )?Sur les? .{1,30} (branche|moyen)s? ((de cassation)|du?e?s? pourvois?)$", None),
H, ALL_PARAGRAPH_AS_TITLE, r"(?i)Consid[ée]rant ce qui suit|soutien(nen)?t que :$|Vu la proc[ée]dure suivante :|^.{1,3} Considérant", None),
TITLE, r"(?i).{0,5}\s*Appr[ée]ciation (de la cour|du tribunal)", None), # CEDH
TITLE, r"(?i).{0,5}\s*(En droit|En fait|Introduction|Conclusion).{0,5}$", None),
TITLE, r"^GRIEFS?$", None), # CEDH
```

For the remaining 45 %. it's more difficult...

Facts

Par courrier daté du 26 septembre 2014, et reçu le même jour au Greffe de la Cour d' Appel de METZ, Monsieur A X a interjeté **appel** de la décision rendue le 19 septembre 2014 par le juge des libertés et de la **détention** de METZ qui a rejeté sa demande tendant sur le fondement de l'article [L.3211-12](#) du **code** de la **santé publique** à obtenir la main levée de ses soins psychiatriques contraints ;

Grounds

Aux termes des dispositions de l'article [L.3212-1](#) du Code de la Santé Publique, une personne atteinte de troubles mentaux ne peut faire l'objet de soins psychiatriques sur la décision du directeur d'un **établissement** de soins mentionné à l'article L.3222-1 que lorsque les deux conditions suivantes sont réunies :

Ses troubles mentaux rendent impossible son **consentement**;

Son **état** mental impose des soins immédiats assortis soit d'une surveillance médicale constante justifiant une hospitalisation complète, soit d'une surveillance médicale régulière justifiant une prise en **charge** sous la **forme** mentionnée au 2° du I de l'article L.3211-2-1.

How would an algorithm
automatically identify legal
sections ?

Plan

About Doctrine

The search challenge

Problem statement

Methods

Results

Conclusions & future work





To complete this task, a human being would take advantage of:

1. The vocabulary used
2. The order of the paragraphs

1. The vocabulary used

Not always so obvious , legislation references in both...

Facts

Par courrier daté du 26 septembre 2014, et reçu le même jour au Greffe de la Cour d' Appel de METZ, Monsieur A X a interjeté **appel** de la décision rendue le 19 septembre 2014 par le juge des libertés et de la **détention** de METZ qui a rejeté sa demande tendant **sur le fondement de l'article L.3211-12 du code de la santé publique** à obtenir la main levée de ses soins psychiatriques contraints;

Grounds

Aux termes des **dispositions de l'article L.3212-1 du Code de la Santé Publique**, une personne atteinte de troubles mentaux ne peut faire l'objet de soins psychiatriques sur la décision du directeur d'un **établissement** de soins mentionné à **l'article L.3222-1** que lorsque les deux conditions suivantes sont réunies :

Ses troubles mentaux rendent impossible son **consentement**;

Son **état** mental impose des soins immédiats assortis soit d'une surveillance médicale constante justifiant une hospitalisation complète, soit d'une surveillance médicale régulière justifiant une prise en **charge** sous la **forme** mentionnée **au 2° du I de l'article L.3211-2-1.**

- > standard approaches

(BoW - TF- IDF)


encodings **performed**

poorly



1. The order of the paragraphs

- Metadata
- Parties
- Composition of the court
- Facts
- Pleas in law and main arguments
- Grounds
- Operative part of the judgment

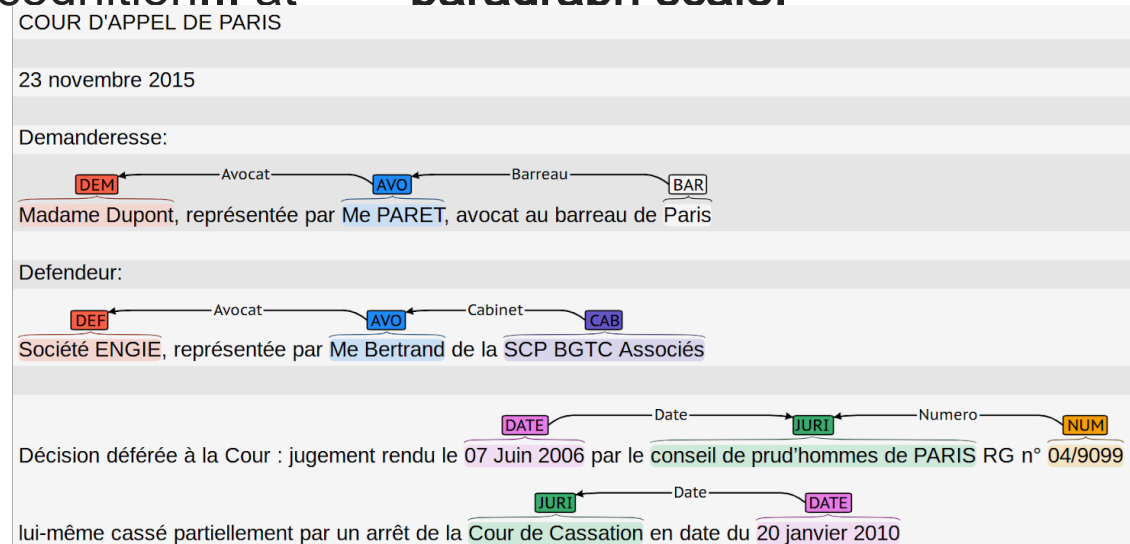


- > sequential
information is
important



Modeling

Looks like Named Entity Recognition... at **paragraph scale**.



With LSTM / CRF, we capture information from

- paragraph inherent properties
- paragraph context (the neighborhood gives insights on the label)

[1] Neural Architectures for Named Entity Recognition. Lample, Ballesteros, Subramanian, Kawakami, Dyer.

NAACL 2016.



~~Exposé du litige:~~

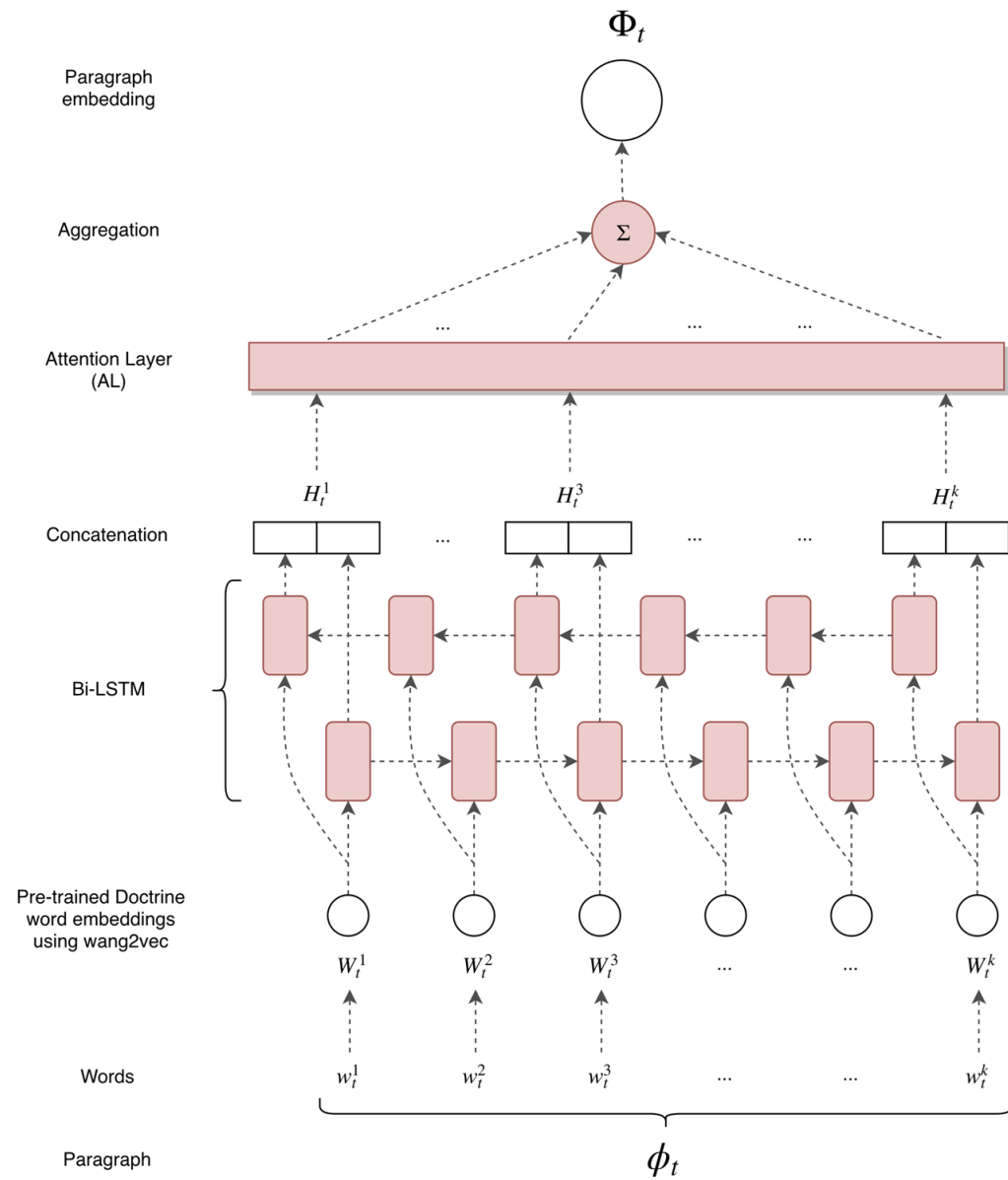
Le 3 septembre 2003, Madame A B, épouse X, se trouvant au marché de XXX et C D, à Y, a buté sur le pied du parasol abritant le stand de fruits et légumes de Monsieur G E F et s'est blessée à l'épaule gauche en tombant au sol.

- Find labeled data from structured decisions with titles
- Remove titles
- Assign each paragraph to its corresponding label (y)
- $y \in [0, 6]$

- > **Supervised classification**



Modeling: paragraph embedding





Modeling: paragraph embedding

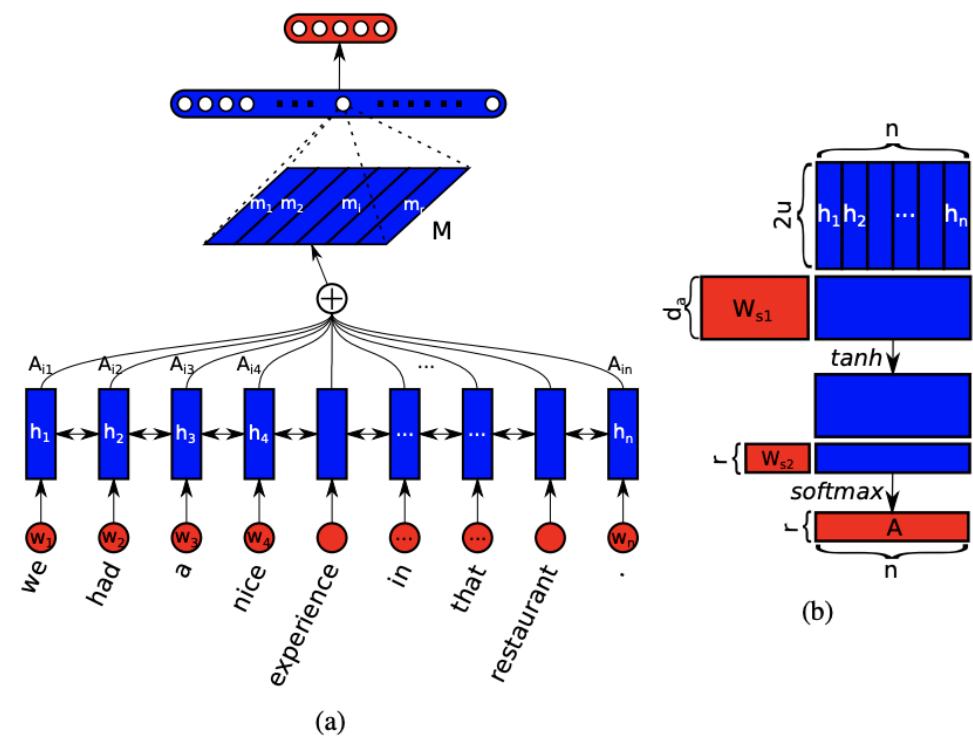
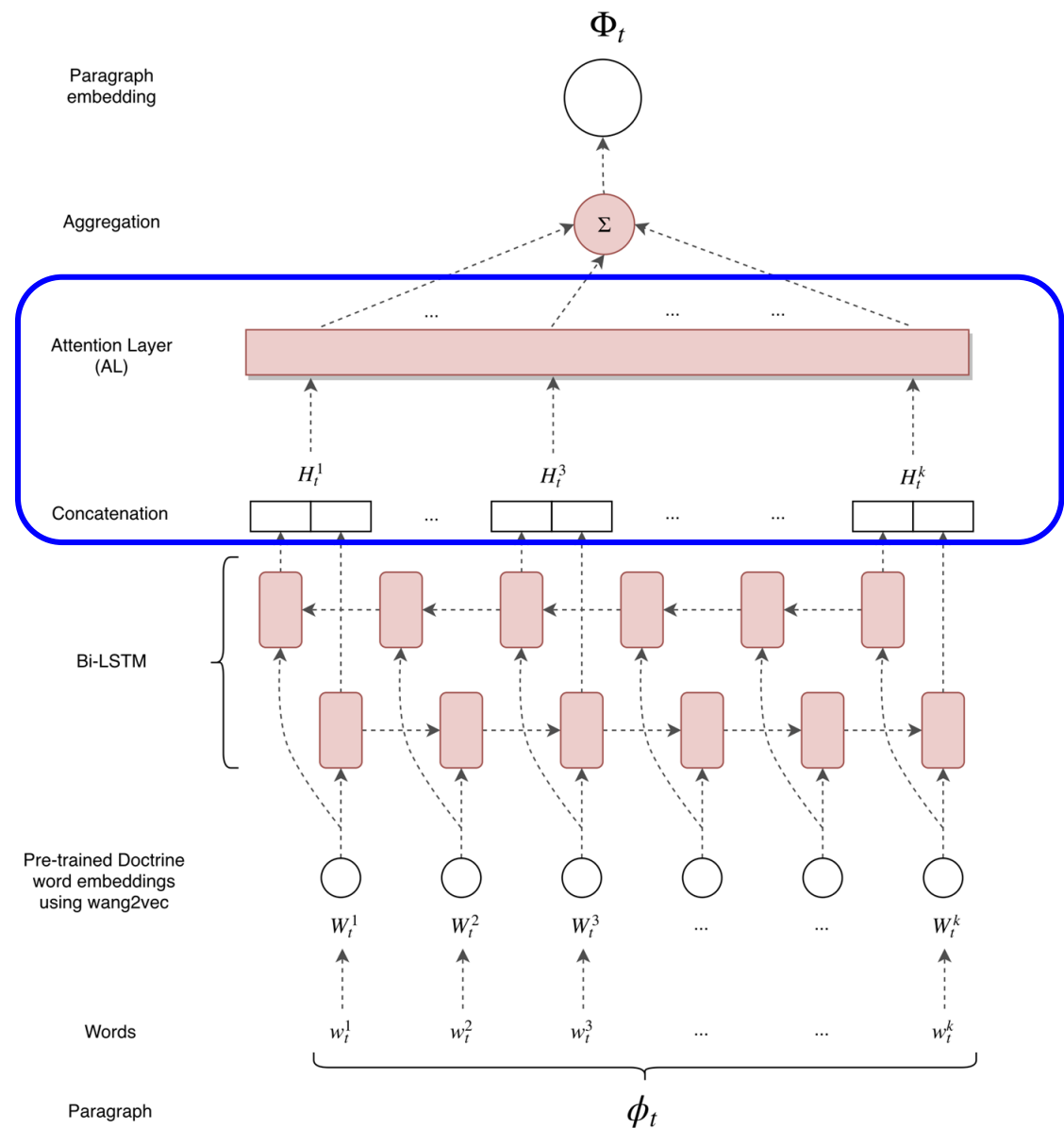
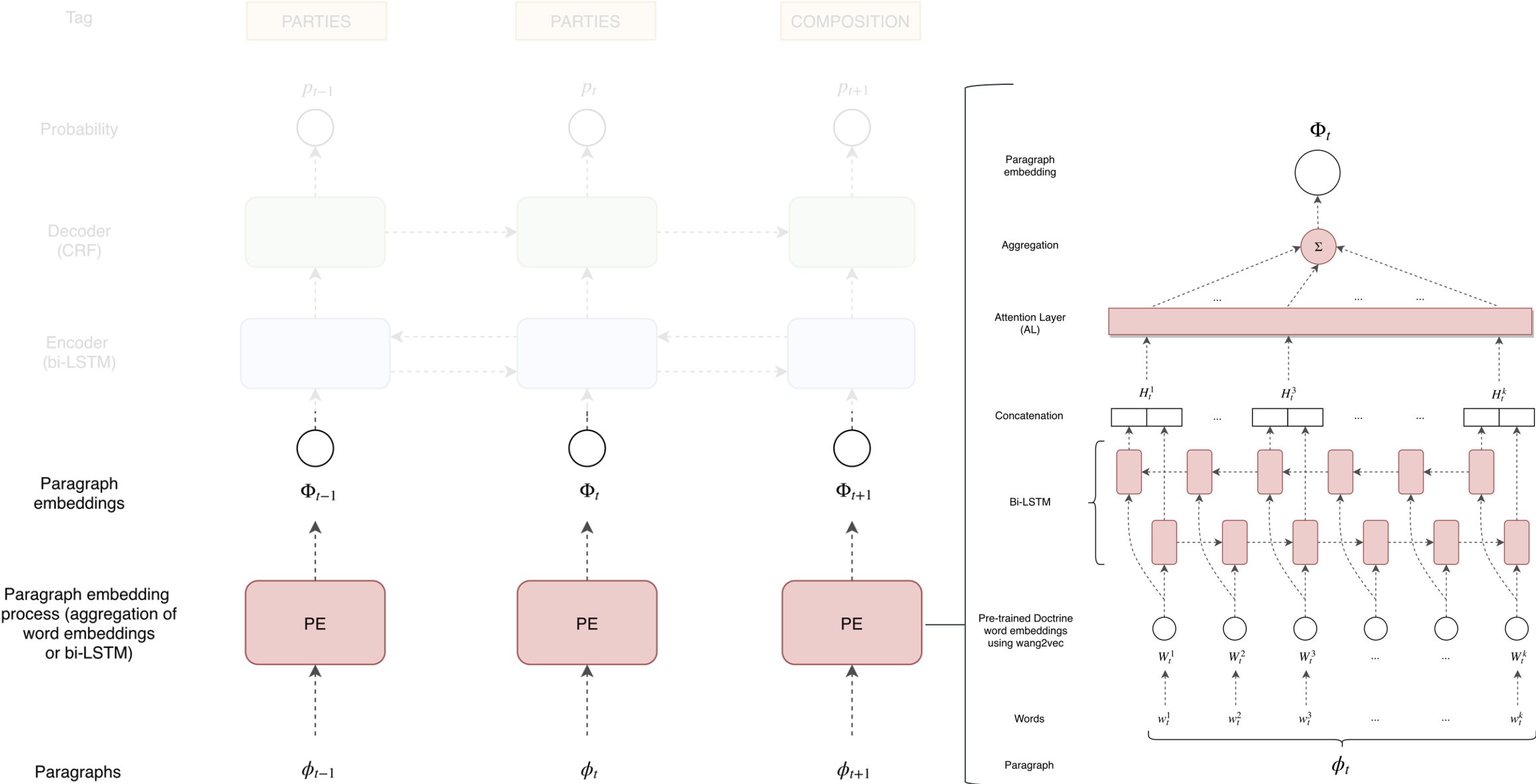


Figure 1: A sample model structure showing the sentence embedding model combined with a fully connected and softmax layer for sentiment analysis (a). The sentence embedding M is computed as multiple weighted sums of hidden states from a bidirectional LSTM (h_1, \dots, h_n), where the summation weights (A_{i1}, \dots, A_{in}) are computed in a way illustrated in (b). Blue colored shapes stand for hidden representations, and red colored shapes stand for weights, annotations, or input/output.

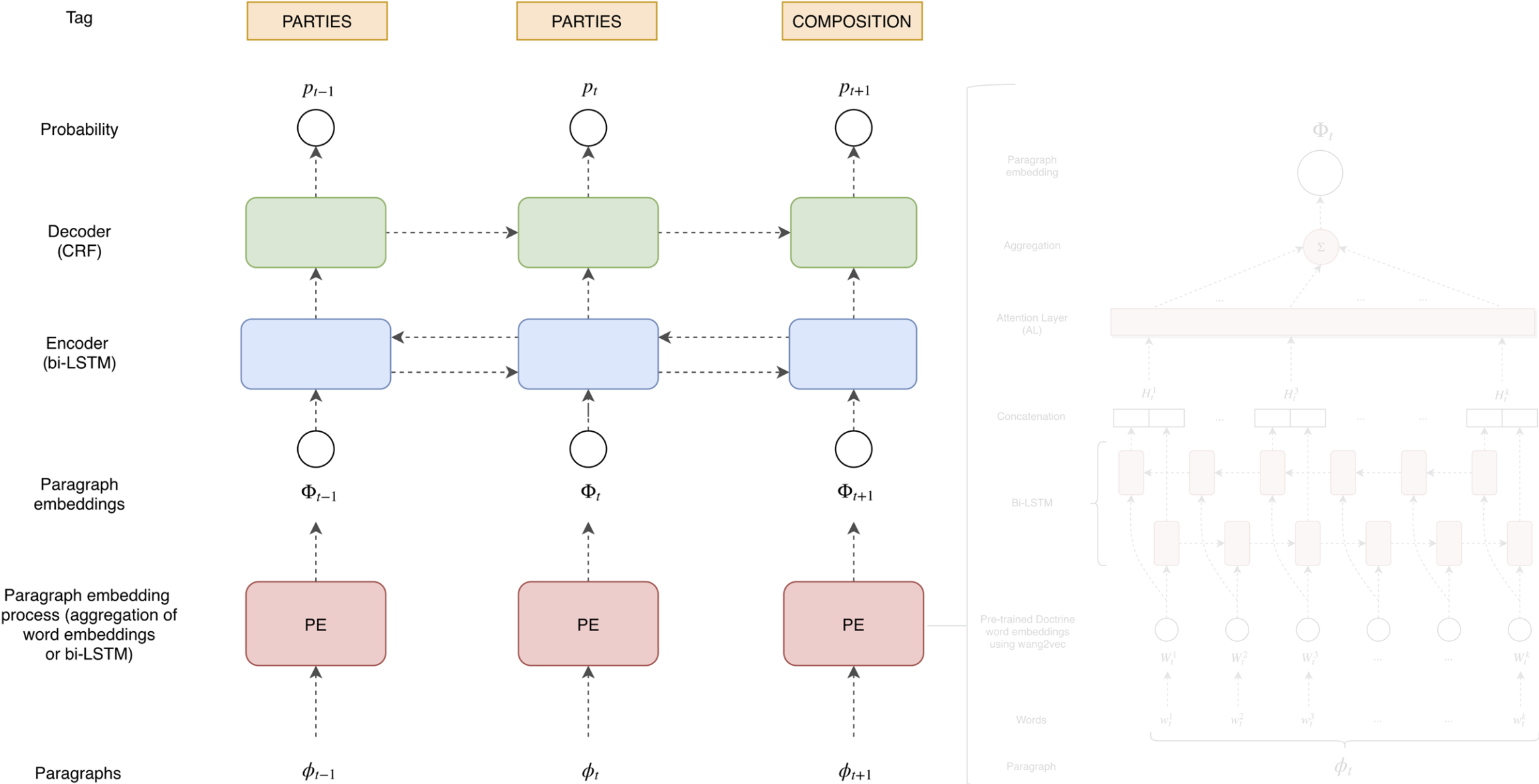


Modeling: entire architecture



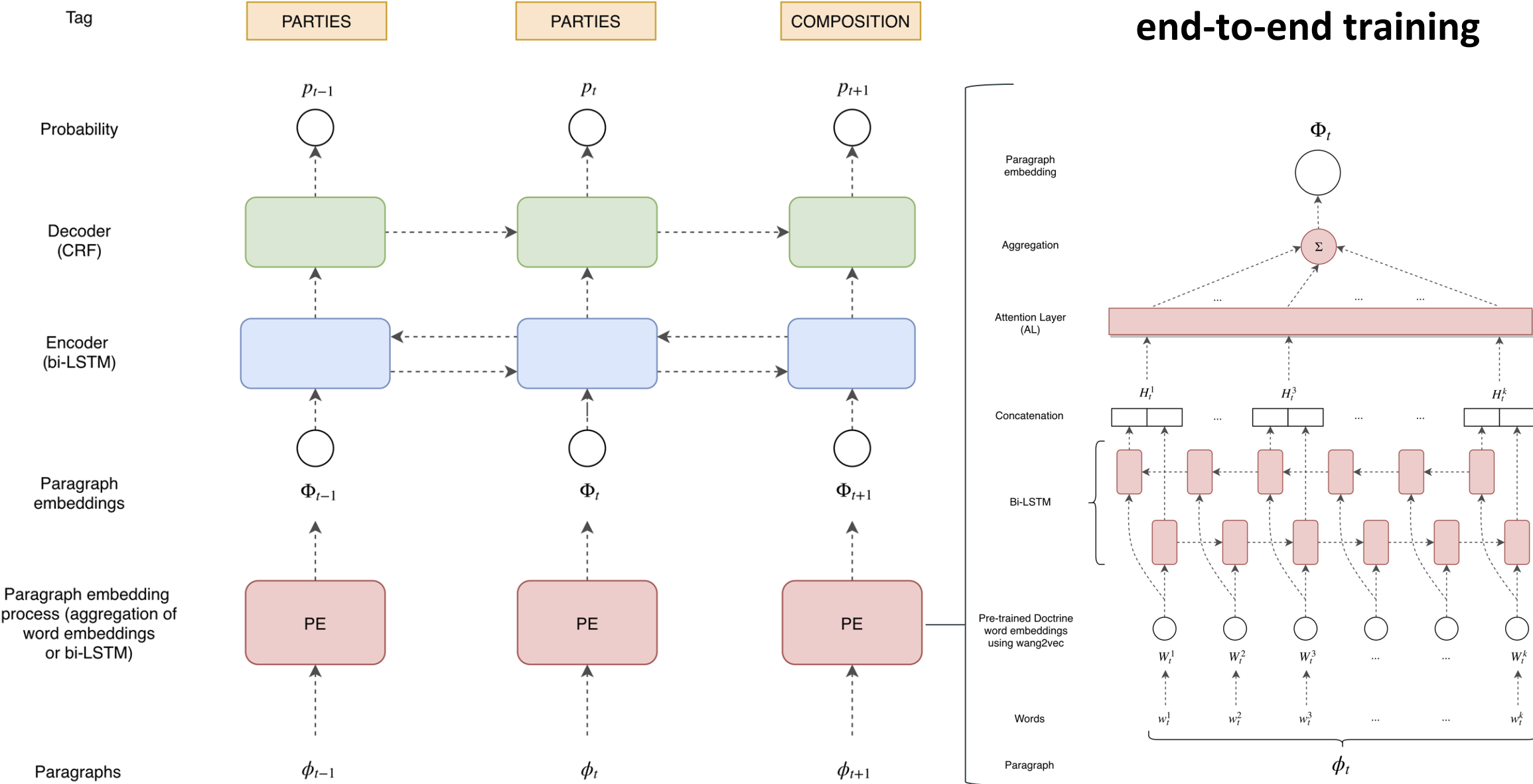


Modeling: entire architecture





Modeling: entire architecture



Plan

About Doctrine

The search challenge

Problem statement

Methods

Results

Conclusions & future work



- Trained on 20.000 decisions

Model	Paragraph embedding	Attention	Decoder	f1-score	Computation (hours)
1	bi-LSTM	Yes	CRF	98.5 %	45
1	bi-LSTM	No	CRF	97.7 %	40
2	bi-LSTM	Yes	softmax	95.9 %	45
2	bi-LSTM	No	softmax	95.8 %	40
3	mean	No	CRF	87.3 %	20
4	mean	No	softmax	86.0 %	20

Table 1: Results of our paragraph classification

Model	Paragraph embedding	Attention	Decoder	f1-score	Computation (hours)
1	bi-LSTM	Yes	CRF	98.5 %	45
1	bi-LSTM	No	CRF	97.7 %	40
2	bi-LSTM	Yes	softmax	95.9 %	45
2	bi-LSTM	No	softmax	95.8 %	40
3	mean	No	CRF	87.3 %	20
4	mean	No	softmax	86.0 %	20

Table 1: Results of our paragraph classification

Model	Paragraph embedding	Attention	Decoder	f1-score	Computation (hours)
1	bi-LSTM	Yes	CRF	98.5 %	45
1	bi-LSTM	No	CRF	97.7 %	40
2	bi-LSTM	Yes	softmax	95.9 %	45
2	bi-LSTM	No	softmax	95.8 %	40
3	mean	No	CRF	87.3 %	20
4	mean	No	softmax	86.0 %	20

Table 1: Results of our paragraph classification

Model	Paragraph embedding	Attention	Decoder	f1-score	Computation (hours)
1	bi-LSTM	Yes	CRF	98.5 %	45
1	bi-LSTM	No	CRF	97.7 %	40
2	bi-LSTM	Yes	softmax	95.9 %	45
2	bi-LSTM	No	softmax	95.8 %	40
3	mean	No	CRF	87.3 %	20
4	mean	No	softmax	86.0 %	20

Table 1: Results of our paragraph classification



On the **45%** incomplete table of contents of Court of Appeal decisions, we now manage to get **90% complete** ones with this approach

ARRET DU 06 MAI 2003 CL/NG ----- 02/00523 -----
S.A. LE MAGICIEN VERT C/ Mickaül F. ----- ARRET N° COUR
D'APPEL D'AGEN CHAMBRE SOCIALE Prononcé à l'audience publique du six Mai
deux mille trois par Catherine LATRABE, Conseillère, La COUR d'APPEL D'AGEN,
CHAMBRE SOCIALE, dans l'affaire ENTRE : S.A. LE MAGICIEN VERT Agropole BP
10 47931 AGEN CEDEX 9 Rep/assistant : M^e LINQUE (avocat au barreau d'AGEN)
APPELANTE d'un jugement du Conseil de Prud'hommes d'AGEN en date du
28 Mars 2002 d'une part, ET : Mickaül F. Rep/assistant : M^e Dominique CAILLIER
(avocat au barreau de TOULOUSE) INTIME :
d'autre part,



- better paragraphs / sentences splitting
- one of the tag is very rare, doesn't perform well
- play with optimizers, dropout, ...
- try different architectures ?

Paragraph classification: [article](#) by Doctrine

18 FÉVRIER 2019



Structuring legal documents with Deep Learning

Facts

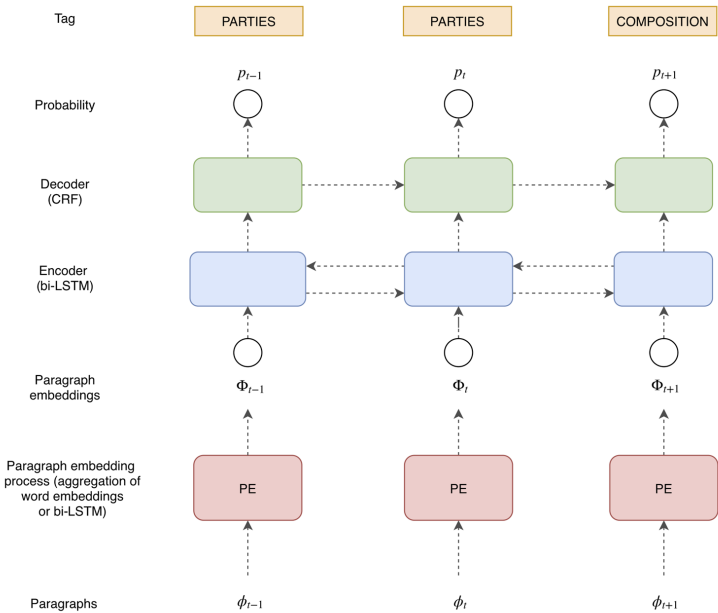
Par courrier daté du 26 septembre 2014, et reçu le même jour au Greffe de la Cour d' Appel de METZ, Monsieur A X a interjeté [appel](#) de la décision rendue le 19 septembre 2014 par le juge des libertés et de la [détention](#) de METZ qui a rejeté sa demande tendant sur le fondement de l'article [L.3211-12](#) du [code](#) de la [santé publique](#) à obtenir la main levée de ses soins psychiatriques contraints;

Grounds

Aux termes des dispositions de l'article [L.3212-1](#) du Code de la Santé Publique, une personne atteinte de troubles mentaux ne peut faire l'objet de soins psychiatriques sur la décision du directeur d'un [établissement](#) de soins mentionné à l'article L.3222-1 que lorsque les deux conditions suivantes sont réunies :

Ses troubles mentaux rendent impossible son [consentement](#);

Son [état](#) mental impose des soins immédiats assortis soit d'une surveillance médicale constante justifiant une hospitalisation complète, soit d'une surveillance médicale régulière justifiant une prise en [charge](#) sous la [forme](#) mentionnée au 2° du I de l'article L.3211-2-1.





Thanks



Additional data

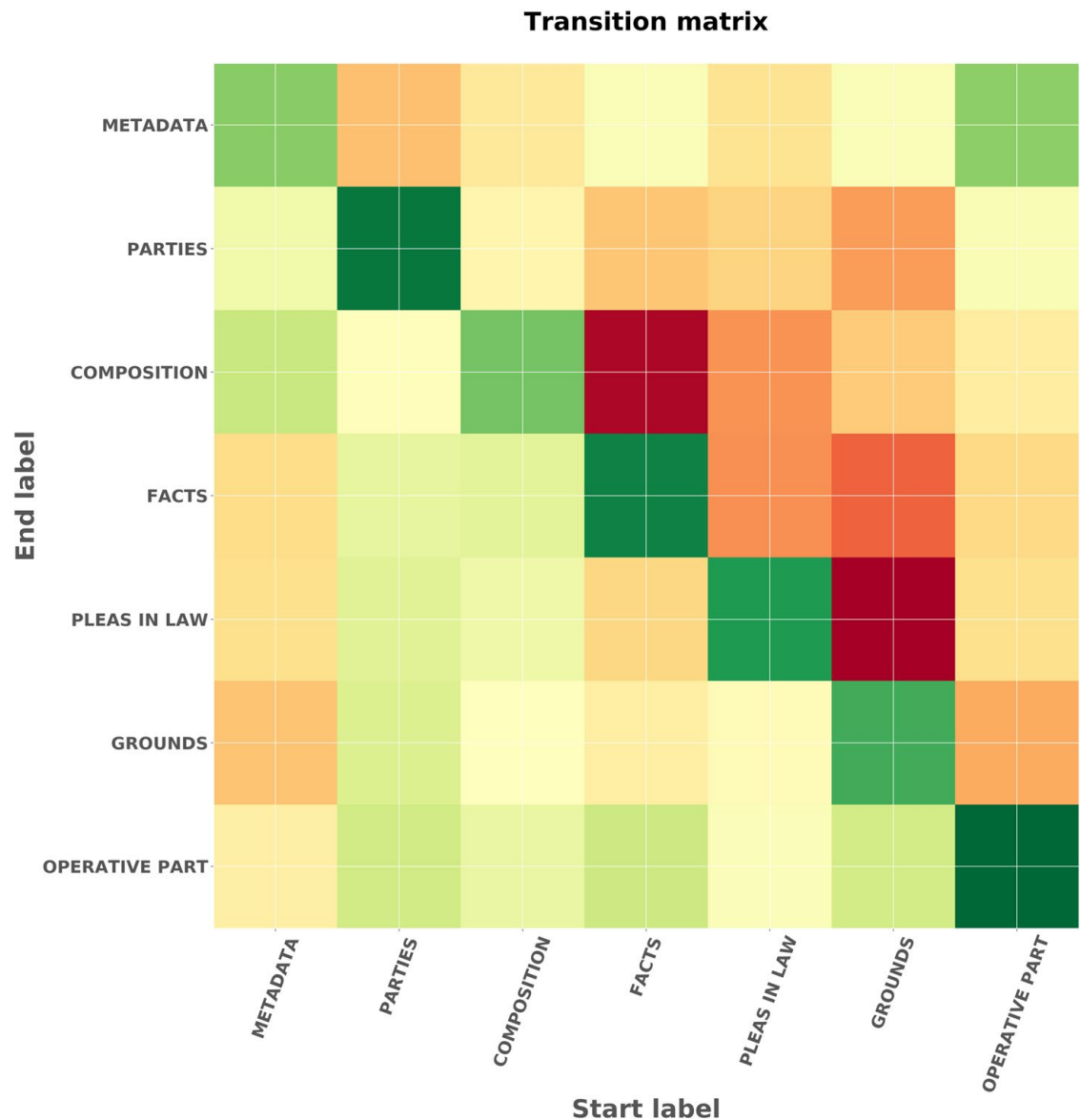




Transition matrix

CRF enables to watch transition
probabilities :

- Each class followed by itself
- Metadata - > Parties
- Metadata - > Composition
- Low triangle part: green
- High triangle part: red



(PARTIES) : assisté de Me Stéphanie SINGER , avocat au barreau de NANTERRE

(COMPOSITION) : Madame Marie @-@ Josèphe JACOMET , président ,

(COMPOSITION) : Monsieur Jean @-@ Loup CARRIERE , conseiller ,

(COMPOSITION) : Madame Anna MANES ,

(COMPOSITION) : qui en ont délibéré ,

(COMPOSITION) : Greffier , lors des débats : Madame Marie @-@ Christine COLLET

(COMPOSITION) : * * * * *

(OPERATIVE) : Confirme le jugement déféré pour le surplus ;

(OPERATIVE) : SE DÉCLARE incompétente ;

(OPERATIVE) : RENVOIE la partie la plus diligente à saisir le juge du fond compétent pour voir statuer sur les modalités d'exercice de l'autorité parentale ;

(OPERATIVE) : RAPPELLE que les mesures provisoires édictées par le juge aux affaires familiales et le conseiller de la mise en état selon ordonnances des de s'appliquer jusqu'à la décision du juge du fond ;

(OPERATIVE) : CONDAMNE A B à payer à C X la somme de 000 euros au titre de l'article 000 du code de procédure civile ,