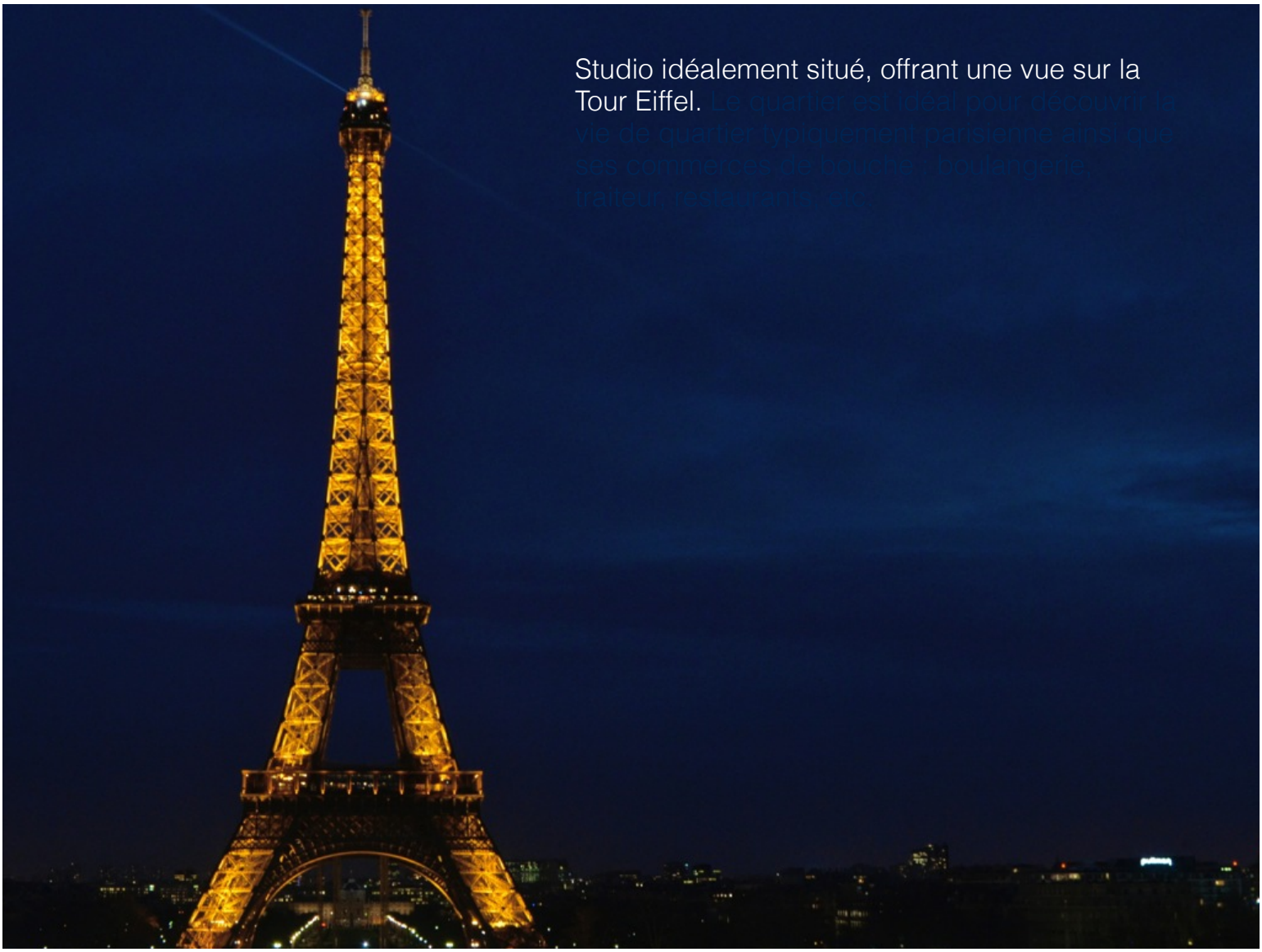
Aspect Term Extraction using Semi-markov CRFs with Word Embeddings

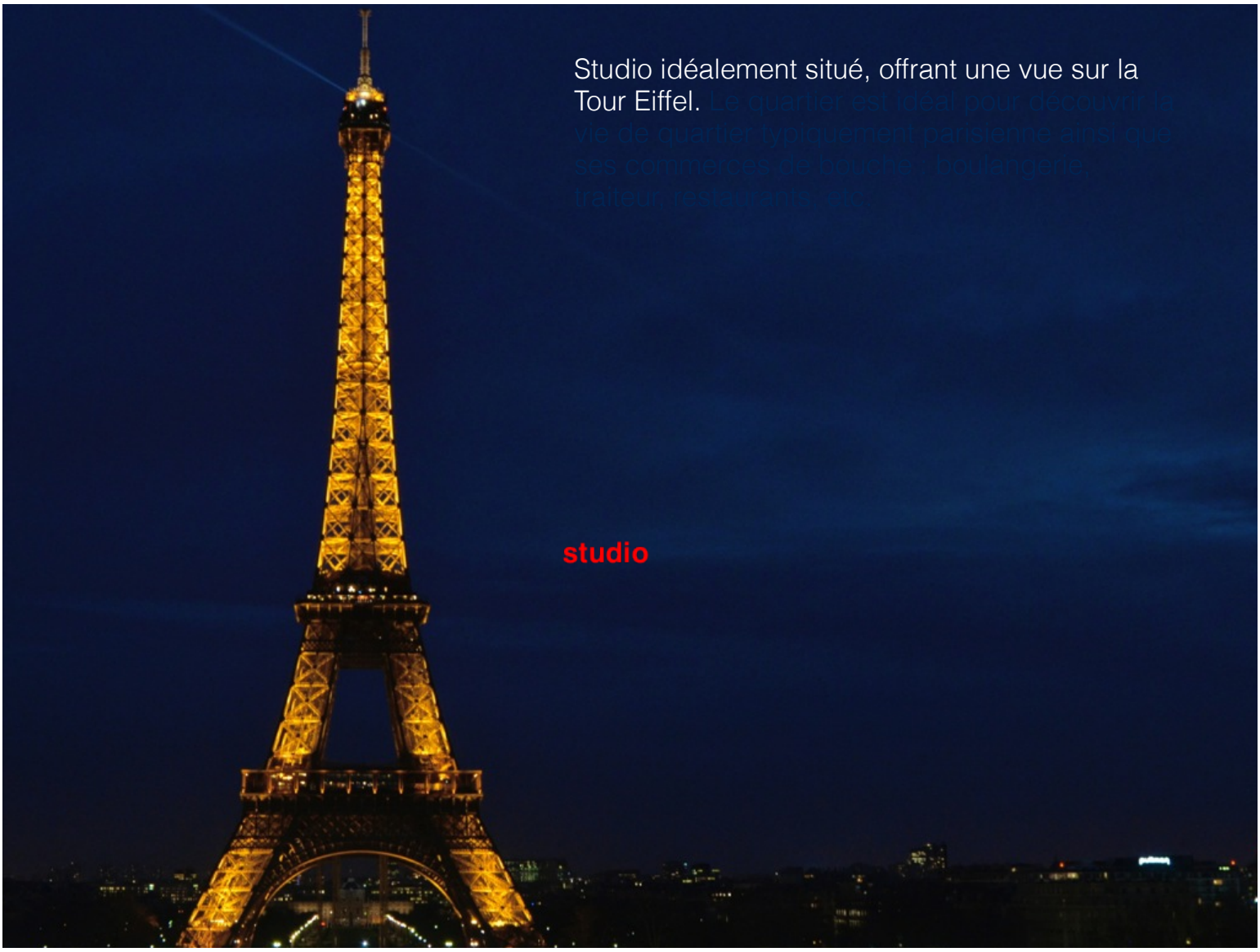
Omer Gunes
omer.gunes@cs.ox.ac.uk

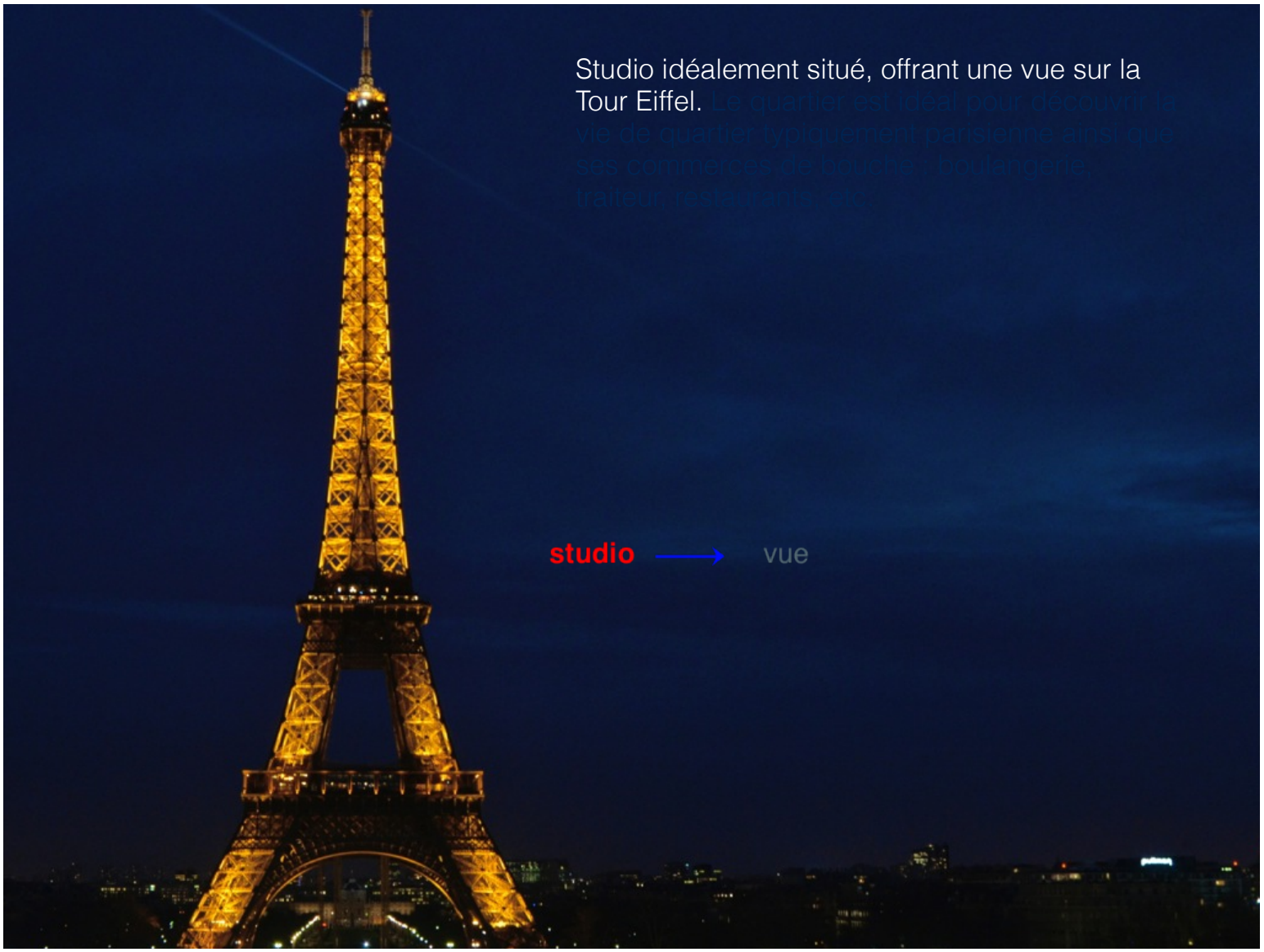
Agenda

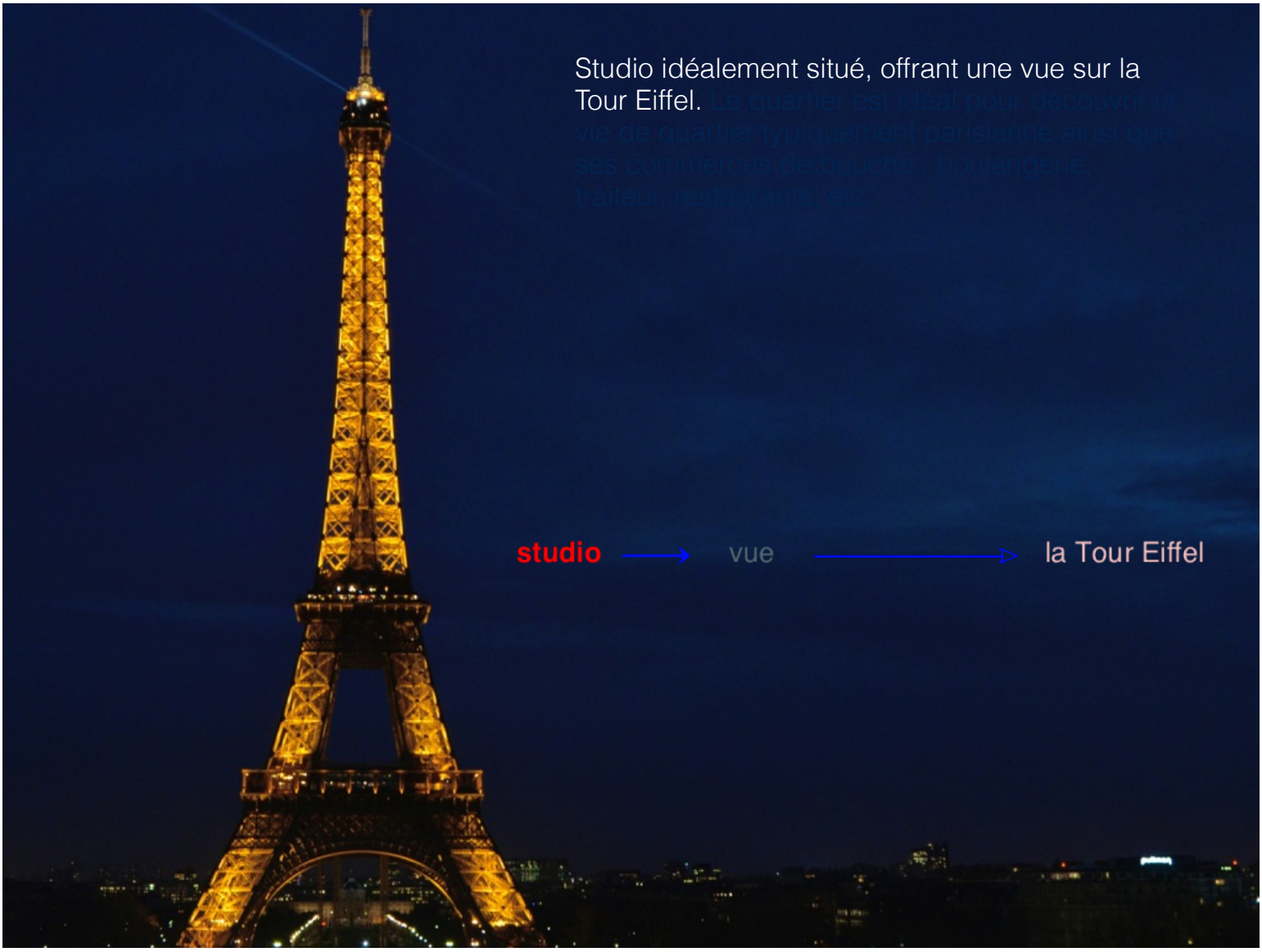
- Aspect Extraction
- Approach
  - Semi-markov CRFs
  - Word Embeddings
- Evaluation
- Related Work
- Conclusion

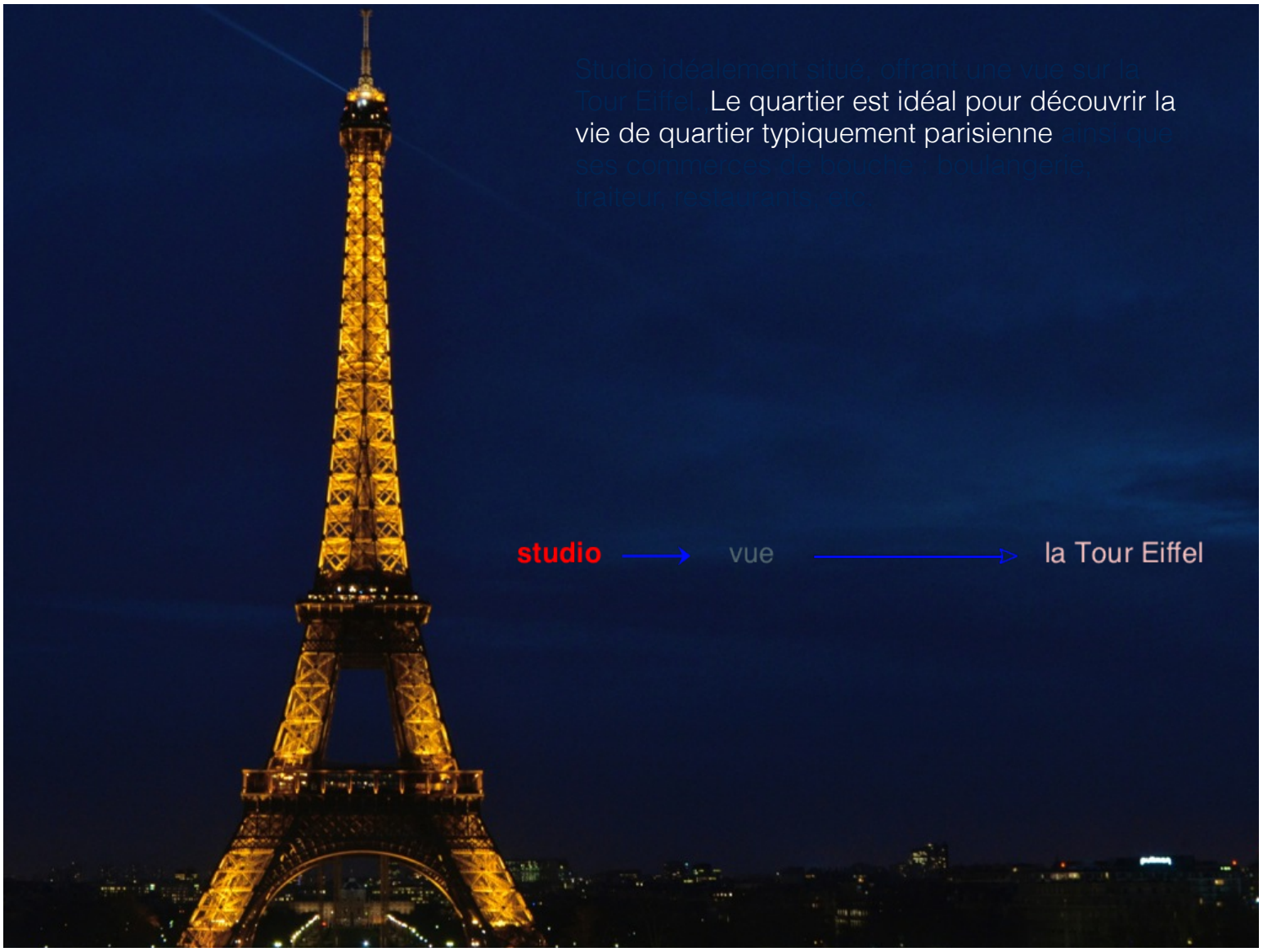


Studio idéalement situé, offrant une vue sur la Tour Eiffel. Le quartier est idéal pour découvrir la vie de quartier typiquement parisienne ainsi que ses commerces de bouche : boulangerie, traiteur, restaurants, etc.


Studio idéalement situé, offrant une vue sur la Tour Eiffel. Le quartier est idéal pour découvrir la vie de quartier typiquement parisienne ainsi que ses commerces de bouche : boulangerie, traiteur, restaurants, etc.
studio

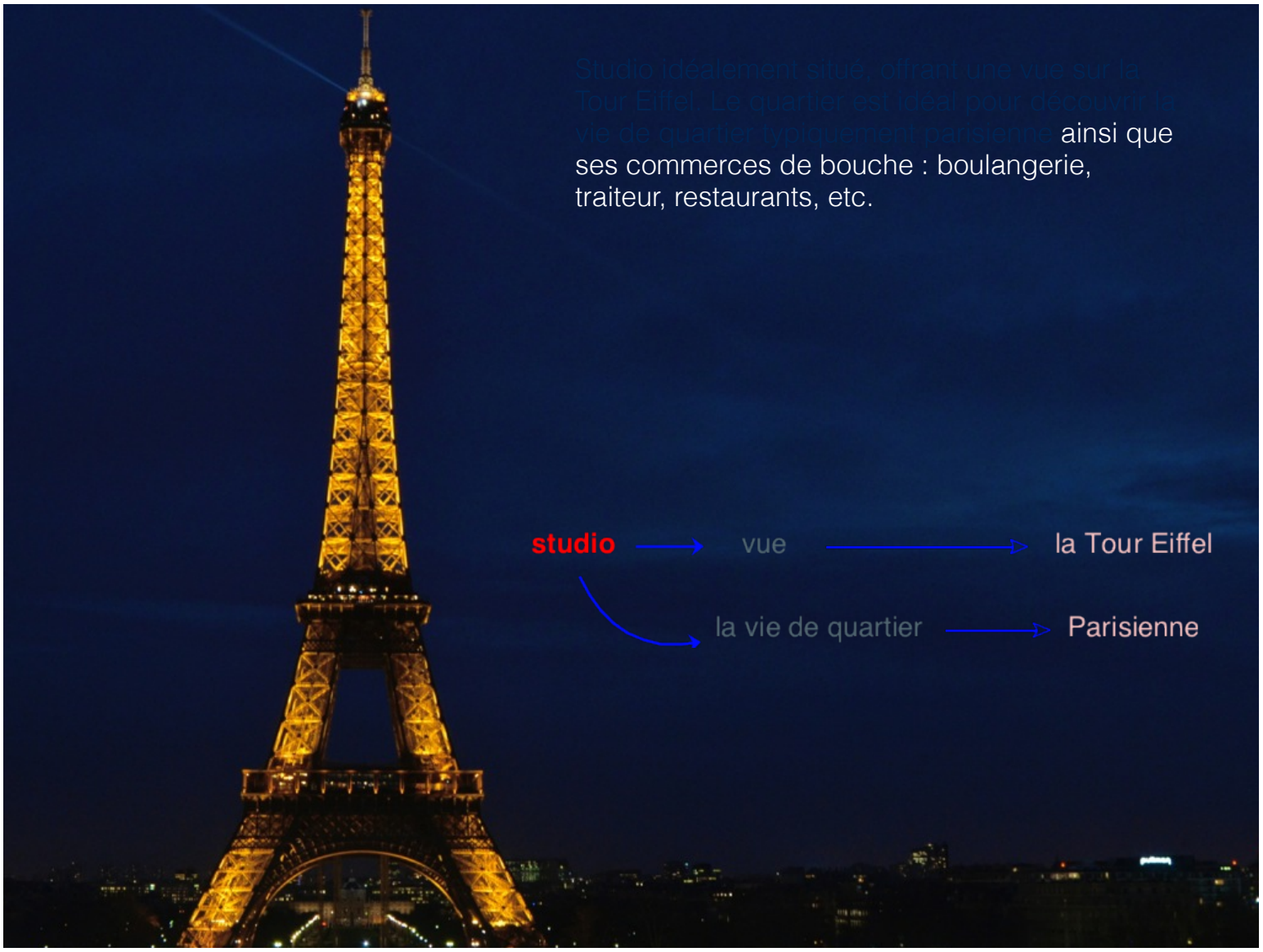

Studio idéalement situé, offrant une vue sur la Tour Eiffel. Le quartier est idéal pour découvrir la vie de quartier typiquement parisienne ainsi que ses commerces de bouche : boulangerie, traiteur, restaurants, etc.
studio
vue


Studio idéalement situé, offrant une vue sur la Tour Eiffel. Le quartier est idéal pour découvrir la vie de quartier typiquement parisienne ainsi que ses commerces de bouche : boulangerie, traiteur, restaurants, etc.
studio
vue
la Tour Eiffel


Studio idéalement situé, offrant une vue sur la Tour Eiffel. Le quartier est idéal pour découvrir la vie de quartier typiquement parisienne ainsi que ses commerces de bouche : boulangerie, traiteur, restaurants, etc.
studio
vue
la Tour Eiffel

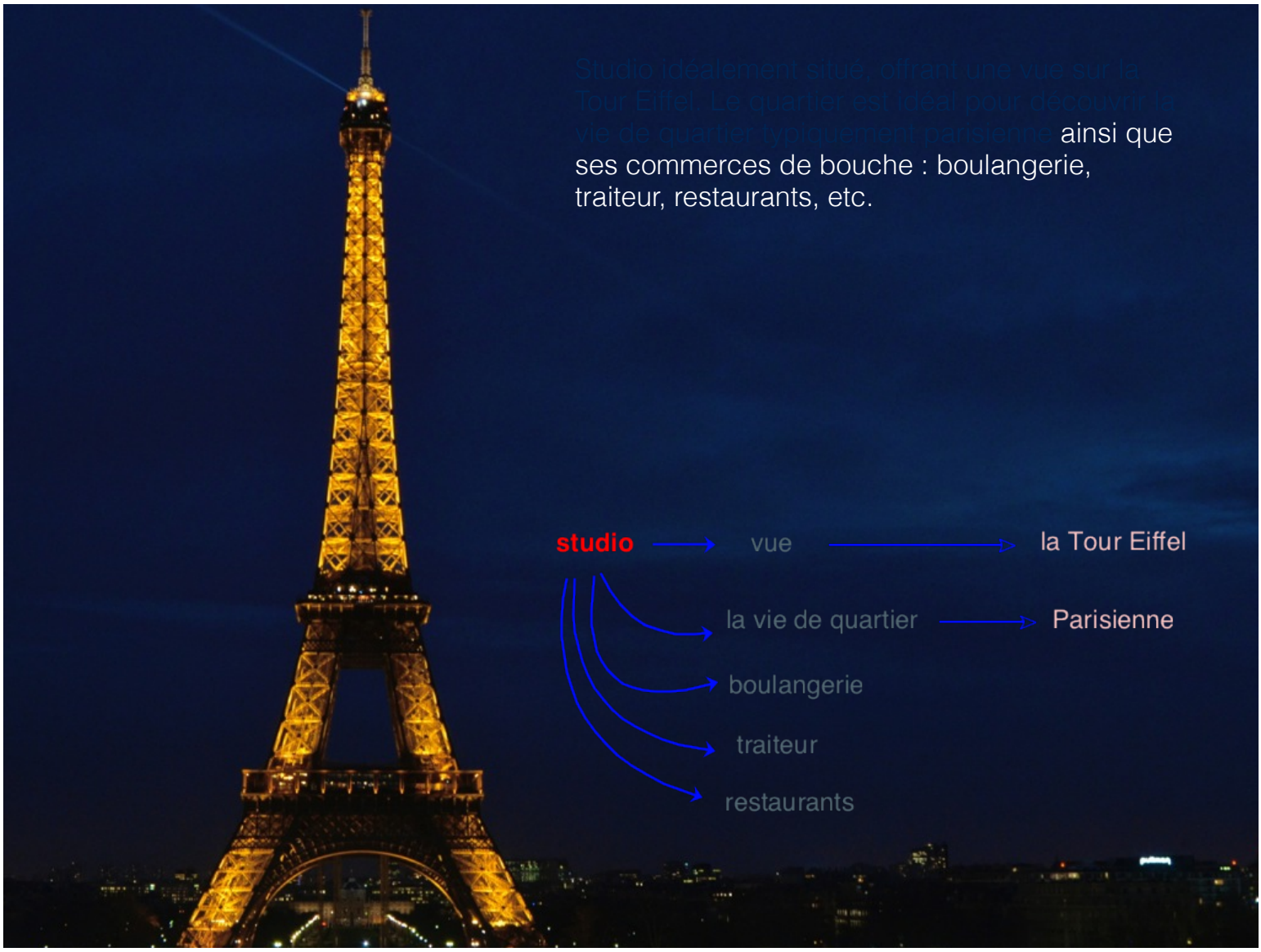

Studio idéalement situé, offrant une vue sur la Tour Eiffel. Le quartier est idéal pour découvrir la vie de quartier typiquement parisienne ainsi que ses commerces de bouche, : boulangerie, traiteur, restaurants, etc.
studio
vue
la Tour Eiffel
la vie de quartier

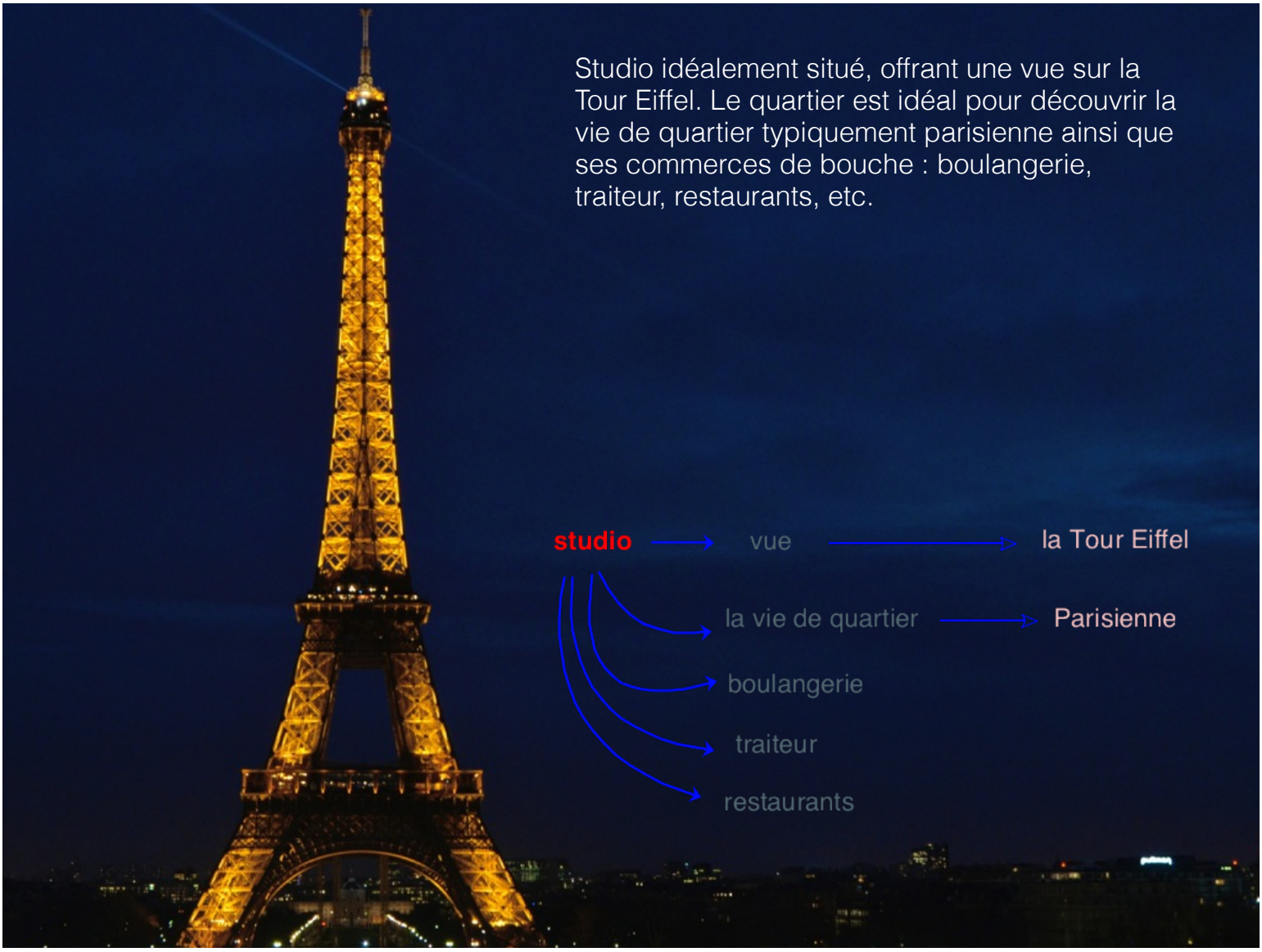

Studio idéalement situé, offrant une vue sur la Tour Eiffel. Le quartier est idéal pour découvrir la vie de quartier typiquement parisienne ainsi que ses commerces de bouche, : boulangerie, traiteur, restaurants, etc.
studio
vue
la Tour Eiffel
la vie de quartier
Parisienne


Studio idéalement situé, offrant une vue sur la Tour Eiffel. Le quartier est idéal pour découvrir la vie de quartier typiquement parisienne ainsi que ses commerces de bouche : boulangerie, traiteur, restaurants, etc.
studio
vue
la Tour Eiffel
la vie de quartier
Parisienne


Studio idéalement situé, offrant une vue sur la Tour Eiffel. Le quartier est idéal pour découvrir la vie de quartier typiquement parisienne ainsi que ses commerces de bouche : boulangerie, traiteur, restaurants, etc.
studio
vue
la Tour Eiffel
la vie de quartier
Parisienne
boulangerie


Studio idéalement situé, offrant une vue sur la Tour Eiffel. Le quartier est idéal pour découvrir la vie de quartier typiquement parisienne ainsi que ses commerces de bouche : boulangerie, traiteur, restaurants, etc.
studio
vue
la Tour Eiffel
la vie de quartier
Parisienne
boulangerie
traiteur


Studio idéalement situé, offrant une vue sur la Tour Eiffel. Le quartier est idéal pour découvrir la vie de quartier typiquement parisienne ainsi que ses commerces de bouche : boulangerie, traiteur, restaurants, etc.
studio
vue
la Tour Eiffel
la vie de quartier
Parisienne
boulangerie
traiteur
restaurants


Studio idéalement situé, offrant une vue sur la Tour Eiffel. Le quartier est idéal pour découvrir la vie de quartier typiquement parisienne ainsi que ses commerces de bouche : boulangerie, traiteur, restaurants, etc.
studio
vue
la Tour Eiffel
la vie de quartier
Parisienne
boulangerie
traiteur
restaurants


Studio idéalement situé, offrant une vue sur la Tour Eiffel. Le quartier est idéal pour découvrir la vie de quartier typiquement parisienne ainsi que ses commerces de bouche : boulangerie, traiteur, restaurants, etc.
Aspect Extractor
studio
vue
la Tour Eiffel
la vie de quartier
Parisienne
boulangerie
traiteur
restaurants

Previous Approaches

- Rule-based
- Minimally-supervised
- Topic modelling-based
- Supervised

Previous Approaches

- Rule-based
- Minimally-supervised
- Topic modelling-based
- Supervised

Previous Approaches

- Rule-based
- Minimally-supervised
- Topic modelling-based
- Supervised
  - CRF-based
  - HMM-based

Approach
Sequence Tagging

“Studio idéalement situé, offrant une vue sur la Tour Eiffel.”

Approach
Sequence Tagging

“Studio idéalement situé, offrant une vue sur la Tour Eiffel.”

- {“studio”, aspect-term, 0, 6}
- {“vue”, aspect-term, 37, 40}
- {“la Tour Eiffel”, aspect-term, 45, 59}

Approach
Semi-markov CRFs

- s: a sequence of consecutive segments
- $\mathbf{s} = \langle s_1, s_2, \ldots s_n \rangle$
  - $s_i = (t_i, u_i, y_i)$
    - t_i: the start position
    - u_i: the end position
    - y_i: the label of the segment s_i

Approach
Semi-markov CRFs

- a feature function $f(x, t_i, u_i, y_i, y_{i-1})$
  - x: the feature
  - t_i: the start position of the current segment
  - u_i: the end position of the current segment
  - y_i: the label of the current segment
  - y_{i-1}: the label of the previous segment

Approach
Semi-markov CRFs

the conditional probability of a segment s given a sequence x

$$p(s \mid x) = \frac{1}{Z(x)} \exp\left(\sum_i \sum_k \lambda_k f_k(x, t_i, u_i, y_i, y_{i-1})\right)$$

where

$$Z(x) = \sum_{s' \in S} \exp\left(\sum_i \sum_k \lambda_k f_k(x, t_i, u_i, y_i, y_{i-1})\right)$$

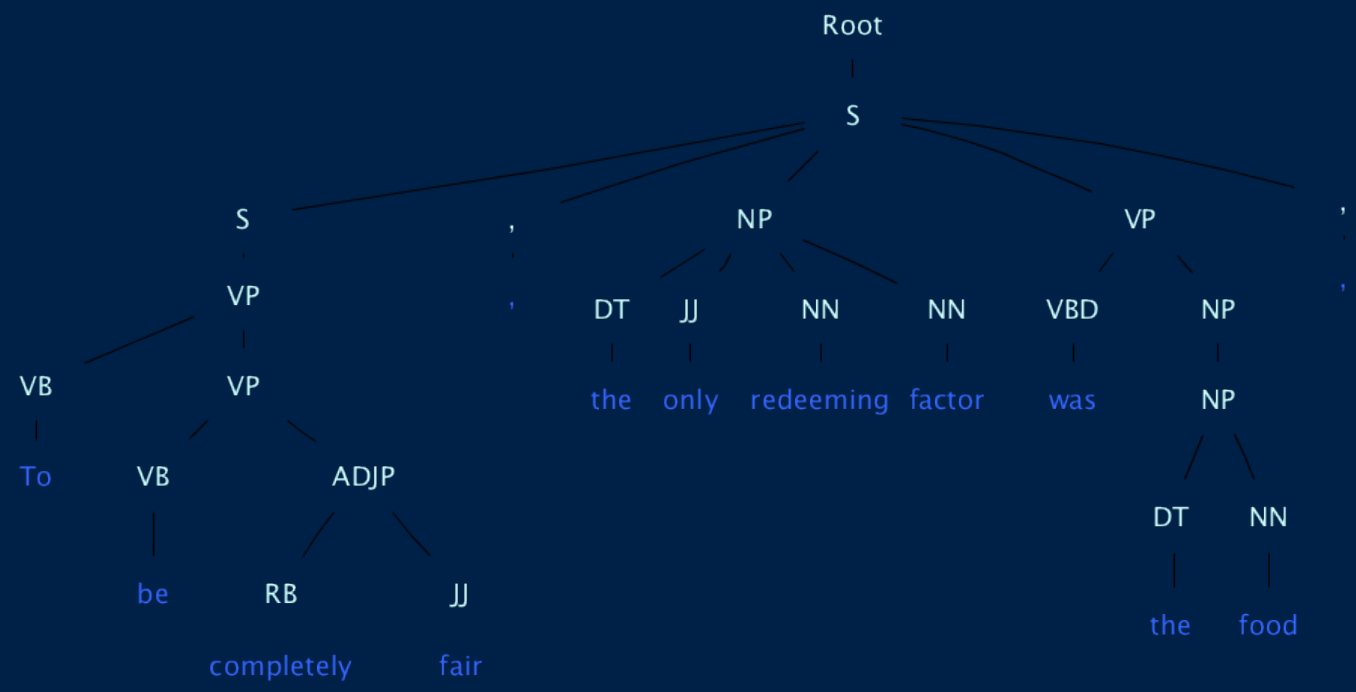
Approach
Features

- CRF-style features
- Segment level feature
- Embedding features

Approach
Features - CRF-style

- the string representation of the word
- the part-of-speech value of the current word
- the orthography value of the current word
- the string representation of the lemma corresponding to the current word
- the noun phrase information of the current word
- a sentiment lexicon entry

Approach
Features - CRF-style - Parsing Information



Approach
Features - Segment Level

- the position of the segment in the current segmentation
- an indicator for the start word within the segment
- an indicator for the end word within the segment
- an indicator before the segment
- an indicator after the segment

Approach
Features - Embeddings

- Brown clustering
  - hierarchical clustering algorithm
- word2vec
  - shallow, 2-layer neural networks
  - predictive
- Glove
  - shallow, 2-layer neural networks
  - count-based

Evaluation

- Datasets
  - SemEval 2014 Task 4
  - SemEval 2015 Task 12
- Metrics
- Results
  - Aspect Term Extraction
  - Opinion Target Extraction

Evaluation Datasets

	ATE (2014)				OTE (2015)						
	Sentences			Aspect Terms	Reviews			Sentences			Targets
Domain	Train	Test	Total	Total	Train	Test	Total	Train	Test	Total	Total
Restaurants	3041	800	3841	4827	254	96	350	1315	685	2000	2499
Laptops	3045	800	3845	3012	277	173	450	1739	761	2500	-
Hotels	-	-	-	-	-	30	30	-	266	266	339
Total	6086	1600	7686	7839	531	299	830	3041	800	3841	2838

Evaluation
Datasets

```
<sentence id="1634">
    <text>The food is uniformly exceptional, with a very capable kitchen which will proudly whip up whatever you feel like eating, whether it's on the menu or not.</text>
    <aspectTerms>
      <aspectTerm term="food" polarity="positive" from="4" to="8"/>
      <aspectTerm term="kitchen" polarity="positive" from="55" to="62"/>
      <aspectTerm term="menu" polarity="neutral" from="141" to="145"/>
    </aspectTerms>
  </sentence>
```

Evaluation Datasets

	ATE (2014)				OTE (2015)						
	Sentences			Aspect Terms	Reviews			Sentences			Targets
Domain	Train	Test	Total	Total	Train	Test	Total	Train	Test	Total	Total
Restaurants	3041	800	3841	4827	254	96	350	1315	685	2000	2499
Laptops	3045	800	3845	3012	277	173	450	1739	761	2500	-
Hotels	-	-	-	-	-	30	30	-	266	266	339
Total	6086	1600	7686	7839	531	299	830	3041	800	3841	2838

Evaluation Datasets

```
<sentence id="1004293:3">
    <text>The food was lousy - too sweet or too salty and the portions tiny.</text>
    <Opinions>
        <Opinion target="food" category="FOOD#QUALITY"
            polarity="negative" from="4" to="8"/>
        <Opinion target="portions" category="FOOD#STYLE_OPTIONS"
            polarity="negative" from="52" to="60"/>
    </Opinions>
</sentence>
```

Evaluation
Metrics

- Two types of setting
  - constrained: only dataset + lexicons
  - unconstrained

Evaluation
Metrics

- Two types of setting
  - constrained: only dataset + lexicons
  - unconstrained
- S: the set of aspect terms/opinion targets
- G: the set of gold-standard annotations

$$P = \frac{|S \cap U|}{|S|}, R = \frac{|S \cap U|}{|G|}$$

Evaluation Results

	ATE (2014)						OTE (2015)					
	Laptops			Restaurants			Hotels			Restaurants		
Domain	P	R	F-1	P	R	F-1	P	R	F-1	P	R	F-1
CRF	86.77	64.81	75.79	88.02	76.24	82.13	73.33	30.84	52.08	73.69	57.27	65.48
CRF+WE	85.33	63.73	74.53	86.54	77.57	82.05	73.91	31.77	52.84	73.13	58.19	65.75
S-CRF	84.96	69.75	77.35	86.19	79.69	82.94	71.68	37.85	54.76	70.27	65.74	68.01
S-CRF+WE	84.01	69.75	76.78	84.76	81.38	83.07	71.18	39.25	55.21	68.69	66.66	67.67

Evaluation Results

	ATE (2014)						OTE (2015)					
	Laptops			Restaurants			Hotels			Restaurants		
Domain	P	R	F-1	P	R	F-1	P	R	F-1	P	R	F-1
CRF	86.77	64.81	75.79	88.02	76.24	82.13	73.33	30.84	52.08	73.69	57.27	65.48
CRF+WE	85.33	63.73	74.53	86.54	77.57	82.05	73.91	31.77	52.84	73.13	58.19	65.75
S-CRF	84.96	69.75	77.35	86.19	79.69	82.94	71.68	37.85	54.76	70.27	65.74	68.01
S-CRF+WE	84.01	69.75	76.78	84.76	81.38	83.07	71.18	39.25	55.21	68.69	66.66	67.67

Related Work

- Aspect Term Extraction
  - Supervised
    - CRF
    - Non-CRF
  - Topic Modelling
  - Embeddings
- Opinion Target Extraction
  - Embeddings

Related Work
Aspect Term Extraction - CRF-based

[Chernyshevich, SemEval, 2014]:

- Instead of inside-outside-begin (IOB)
  - *FA:* preceding attribute word of the head noun
  - *FPA:* succeeding attribute word of the head noun
  - *FH:* head noun
  - *O:* non-aspect

Related Work
Aspect Term Extraction - CRF-based

[Toh et Wang, SemEval, 2015]:

- External sources
  - *WordNet Taxonomy*
  - *Word Clusters*
  - *Name Lists*

Related Work
Aspect Term Extraction - non-CRF-based

[Kiritchenko, SemEval, 2014]:

- Semi-markov HMMs
- Instead of inside-outside-begin (IOB)
  - *outside*
  - *problem*
  - *treatment*
  - *test*
- Up to 30 tokens except outside

Conclusion

- Semi-markov CRFs for ATE
  - Better results compared to CRFs
- Word embeddings as features
- Best scores for each domain
  - SemEval 2014 Task 4
  - SemEval 2015 Task 12

Demo

http://www.oxtractor.com