Student: Cécile Grivaz
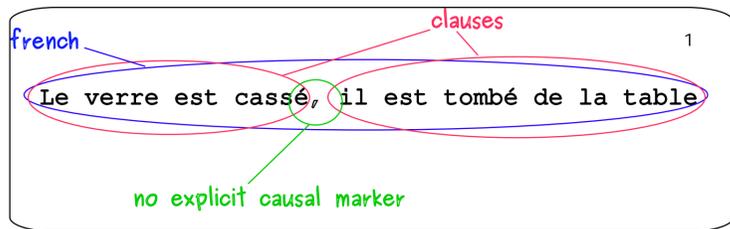Advisors: Jacques Moeschler and Martin Rajman

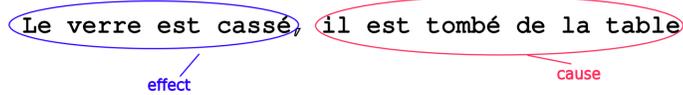# A project for automatic causal knowledge extraction from clauses in French

Goal

We want to develop a computer program which, when given French text, is able to find pairs of clauses in a causal relation, even if there is no causal marker, such as a connector.

french — clauses — 1
Le verre est cassé, il est tombé de la table
no explicit causal marker

INPUT → OUTPUT

Is there a causal relation ? yes

Where are the cause and the effect expressed ?

Le verre est cassé, il est tombé de la table
effect — cause

Generalisation

tomber(verre: v1, table) **CAUSE** se casser(v1)

Methodology

Using an incremental process and easy to find examples, we want to learn what the properties of causation are in order to be able to find more difficult examples. We make the hypothesis that explicit examples display the same features as implicit ones.

Bootstrap
First, we need some positive examples in order to learn the features of causation. We use easy-to-find causal sentences because they contain explicit markers. We can retrieve the examples from a big corpus using lexico-syntactic patterns such as VP parce que VP. [2]

Nous sommes le 24 décembre et chacun ouvre ses cadeaux. Je suis heureux parce que j'ai reçu un grille pain. Je vais pouvoir me faire un magnifique petit déjeuner. [3]

Match VP parce que VP

Corpus of positive examples
…
Je suis heureux parce que j'ai reçu un grille-pain.
…

Incremental process
More positive examples allow us to create a more precise and complete model of common causes and effects. With more positive examples, we can compare new candidates to more precise types of causes and effects, thereby allowing more structural variation.

Learn common causes and effects
Store in a database, information on what types of event cause other types of event. This information can be gained from previous positive examples.

Je suis heureux parce que j'ai reçu un grille-pain.

get propositions

recevoir(humain:h1, grille-pain) **CAUSE** être_heureux(h1) [5]

generalize

obtenir(humain:h1, appareil électro-ménager) **CAUSE** sentiment_positif(h1) [6]

Learn structure
Use a machine learning algorithm to learn from positive examples. The feaures should come from theoretical work. For example: verb tenses, aspectual classes, …

Je suis heureux parce que j'ai reçu un grille-pain.
present   state   passé composé   event

New candidate
This is an example of a sentence we want to analyze which the system has never seen before:

Jean est content, il a acheté un mixeur. [4]

Analyze new candidate
Extract structure:

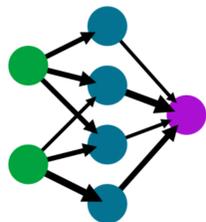**Jean est content, il a acheté un mixeur.**
present   state   passé composé   event

Get propositions and generalize

obtenir(humain:h1, appareil électro-ménager) **?** sentiment_positif(h1)

Use a machine learning algorithm to compare new candidates to previous positive examples and decide if the candidate is causal.

Neural Network

Learning Examples

Positve examples:
explicitly causal sentences extracted from a large corpus.

Negative examples:
- Any other sentences from the corpus ? Risk of including positive but not explicit causal examples (the very sentences that we want to learn how to extract).
- Other sentences that are explicitly not causal ? For example, sentences formed of two clauses separated by the non causal marker "mais"/"but". Risk of not including sentences that are not causal but which lack a non causal marker. This could mean that the negative examples are a sub-space of the class we want to learn to discriminate against.
- Ad hoc method for learning only from positive examples ?

Features

The features should be chosen from lingustic theoretical work and manual examination of a corpus. They fall into two categories: structural features and world knowledge about common cause and effects.

Structural features:
Typically, these features consist of several classes, e.g. the tense of the main verb of the first clause may be present, future, etc. Those classes should be coded as several features with a binary value (present/ not present). For example, a main verb in future tense would make the following features: present: 0, future:1, passé composé:0, ….

Common cause and effect features:
This feature will probably be a float. We need to develop a measure that takes into account the probability that the two verbs of the candidate relate to two classes of cause and effects verbs in the database, as well as the probability there is a causal relation between verbs of these classes.

Translations:

1: The glass is broken, it fell from the table.
2: parce que: "because"
3: Today is the 24th of December and everybody is opening gifts. I am happy because I got a toaster. I will be able to make myself a beautiful breakfast.
4: John is glad, he bought a blender.
5: receive(human:h1, toaster) CAUSE be_happy(h1)
6: get(human:h1, equipment) CAUSE positive_feeling(h1)

Open Questions

-How to learn from only positive examples ?
-How best to handle a mix of binary and float features ?