

1. Description

The syntactic structure of sentence can be described as a rooted tree that indicates the syntactic relationships between the words of a sentence a in Fig. 1. In these trees, there is a particular vertex, called root, that has no incoming edges. The backbone of the tree is the free tree (Fig. 1 (c)), that is the undirected tree that results from removing link direction from the rooted tree (Fig. 1 (b)).

There has been a lot of research on extracting those trees automatically from texts using unsupervised methods. These methods are critical in contexts where there is no training dataset because we know very little about that language (e.g., in low resourced languages or languages that deviate from the frame of most Indo-European languages). However, a serious limitation of these methods is that they make a lot of mistakes concerning the direction of the arcs. These methods often guess correctly that two words x and y are linked but they fail to guess if $x \rightarrow y$ or $x \leftarrow y$. There are distinct ways the right direction of the arcs can be guessed. One is by identifying the root in the free tree and then assigning arc direction consistently from that root. Notice that in a rooted tree no two vertices can point to the same vertex.

The aim of this project is to develop supervised methods that predict the root node of a free tree to improve supervised an unsupervised parsers in the future. This project is just on the supervised problem of predicting the root in a simplified setting:

1. An instance is a series of features about a vertex in the free tree. Some examples of possible features are: the degree of the vertex, its average distance to other vertices in the tree.
2. Given that information, the model has to predict whether the vertex is a root or not.

As only information about the rooted tree can be exploited, certain features cannot be used:

1. Information having to do with the position of the vertex in the sentence, e.g. the position of the vertex in the sentence or the length of its arcs (the length of an arc linking vertices in positions i and j of the sentence is $|i - j|$).

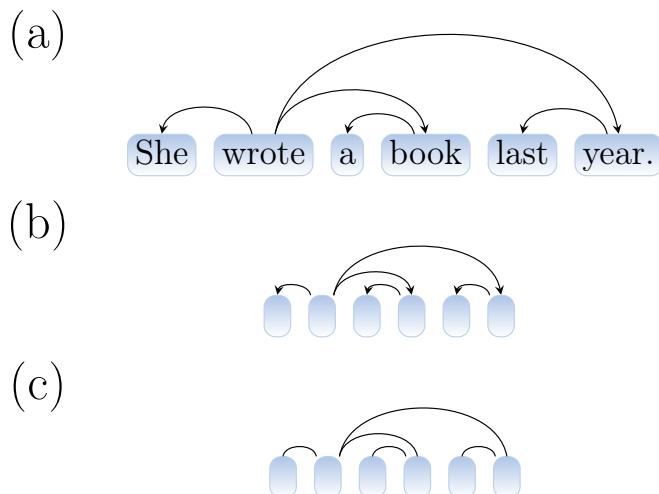


Figura 1: (a) The syntactic dependency structure of a sentence. (b) The corresponding rooted tree. (c) The corresponding free tree.

2. Information about the word that corresponds to the vertex (its string, its part-of-speech,...).
3. The language of the sentence where the vertex appears.
4. As the prediction has to be made on the free tree, information on the rooted tree cannot be used. For instance, you cannot use the in-degree of the vertex. That would make the problem trivial because the root vertex is the vertex that has zero in-degree.

The source data is the Parallel Universal Dependencies (PUD) collection. PUD consists of a series of sentences and their syntactic dependency annotation. You have to use a dataset that contains just the rooted trees of each sentence. The rooted trees (Fig. 1 (b)) are available in two variants, PUD and PSUD, that depend on how the structure of a sentence has been defined. The dataset to be is PUD 2.6 or PSUD 2.6, which are available at <https://cqlab.upc.edu/lal/universal-dependencies/>.

The project has three phases that are to be reflected in the organization of the report.

1.1. Preparation

You have to extract a matrix from the PUD collection to train supervised models. The matrix consists of attributes (vertex features) and additional column that is the class of the vertex. All the features are computed on the free tree (Figure 1 (c)) corresponding to the rooted tree (Figure 1 (c)) that is available in the dataset. The class is 1 (the vertex is a root) or 0 (the vertex is not a root). In that matrix, every row corresponds to a vertex in a sentence. The features to be used are open but subject to the rules explained above. The minimum futures to be used are vertex degree and distance to other vertices in the tree.

Some advanced features are whether the vertex is a center or a centroidal vertex, as defined in Harary's classic book.¹. A gentle definition can be found here https://oeis.org/A000055/a000055_7.pdf. A center (or Jordan center) vertex of a tree is vertex of minimum eccentricity, that is, a vertex u such the greatest distance $d(u, v)$ to other vertices v is minimal. In a tree, there can be one or two centers. A centroidal vertex of a tree is a vertex that minimizes the maximum size of the branches that emanate from any vertex. or centroidal vertex. In a tree, there can be one or two centroidal vertices.

Centers and centroids of a tree (as well as simpler characteristics of a vertex) can be computed easily using the Linear Arrangement Library.

You may need to apply some transformation of the data, e.g. rescaling or normalization.

You have to deliver that matrix. It is possible to use more than one matrix.

1.2. Modelling

You have to apply methods that we have seen in the classroom and can use any of the methods available at **scikit-learn**. The evaluation must be based on at least the following scores: precision, recall and an F -measure (the harmonic mean of precision and recall). You have to consider at least the following baselines:

1. A model that always predicts if the vertex is not a root. Motivation: only one vertex of a sentence can be a root.
2. A model that predicts that the vertex is a root at random (with probability p), namely by flipping a coin where sides may not be equally likely.

¹(1969). Graph theory. pp. 35-36 [https://users.metu.edu.tr/aldoks/341/Book%20\(Harary\).pdf](https://users.metu.edu.tr/aldoks/341/Book%20(Harary).pdf)

1.3. Research and discussion

Some research questions:

1. Which models are the best? Which features are more predictive?
2. In which cases the baselines perform better?
3. What is the accuracy of the models depending on the language? Remember that you cannot use the language as a feature for training.
4. What is the accuracy of the models depending on the length of the sentence?

2. Entrega

Tendréis que entregar a través del Racó (pestaña “Prácticas”) vuestra solución a la práctica antes de la fecha indicada en el mismo Racó.

La práctica es por parejas. Solamente hace falta que un miembro entregue la solución por el Racó, eso sí, indicad claramente los nombres de las personas autoras del proyecto en el informe. La entrega deberá contener

1. El código que hayáis utilizado
 2. La matriz o matrices de datos que hayáis generado para el entrenamiento de modelo.
 3. Un informe (breve, 4-5 páginas) detallando qué metodología habéis seguido, resultados, dificultades que os habéis encontrado y cualquier cosa que penséis que haya sido relevante a la hora de realizar el proyecto. Si necesitáis más espacio, podéis usar apéndices para la información complementaria.
- Debéis indicar claramente la fuente usada, PUD o PSUD, y definir claramente el significado de cada atributo elegido.

Algunas personas deciden no formar una pareja y trabajar en solitario. Los motivos son variados:

- No han encontrado pareja. Respuesta: si los alumnos no han establecido una canal de comunicación para formar pareja es un problema suyo por el que serán evaluados. No es un problema del profesor. Es habitual que los alumnos que no van a clase tengan dificultades para encontrar pareja, por eso la práctica también evalúa indirectamente la asistencia a clase o el saber suplir adecuadamente la no asistencia a clase. **No es responsabilidad del profesor ayudar a los alumnos a buscar pareja.** En la práctica también se evalúa la capacidad para saber usar las redes sociales. Los alumnos pueden decidir por su cuenta y riesgo no estar conectados con el resto de estudiantes de la clase mediante las redes sociales.
- No saben o no les gusta trabajar en grupo. Respuesta: en la práctica se evalúa la capacidad para trabajar en equipo.
- Trabajan mejor solos. Respuesta: demuestran que deben aprender a trabajar en equipo.
- No pueden trabajar en equipo. Respuesta: falso, incluso se puede colaborar online. Si no tienen tiempo para trabajar en la práctica entonces no pueden contradecirse entregando la práctica.

Algunos alumnos de este tipo escriben al profesor pidiéndole que les deje entregar en solitario o unirse a una pareja para formar un trío con la excusa de que no han conseguido encontrar pareja o porque ya es demasiado tarde encontrarla. Estas excusas no sirven (ver respuestas más arriba). El truco es antiguo. El profesor es viejo. Por eso los siguientes criterios de evaluación. Puede darse que haya un número impar de alumnos que realmente quieran entregar la práctica. **Entonces, la nota de la práctica se calcula de la forma siguiente:**

- Si se trata de una pareja, la nota es el resultado de evaluar el contenido de la práctica.
- Si se trata de un equipo de más de dos personas la nota es un cero directo.
- Si se trata de una estudiante en solitario, la nota es el resultado de evaluar el contenido de la entrega si se trata de la única entrega en solitario. Si hay más de una entrega en solitario, la nota es un cero directo para cada estudiante que entrega en solitario.