

# Sumarização Automática de Texto

Orientador: João Paulo Cordeiro

15 de julho, 2019

## Introdução

Sistemas automáticos capazes de reter o essencial da informação contida num texto, descartando tudo aquilo que é acessório e redundante, são de extrema importância, com implicações óbvias noutras áreas do *Processamento da Linguagem Natural* (PLN) [1] e aplicações muito relevantes nos paradigmas futuros da interação homem-máquina [2, 3]. Todavia, apesar de um contínuo esforço de investigação na área da *Sumarização Automática de Texto* (SAT), desde os primórdios da PLN [4, 5, 6], a qualidade dos resultados produzidos ainda deixa muito a desejar, quer a nível quantitativo (métricas ROUGE [7]) quer a nível de apreciação qualitativa humana. Neste contexto, o presente trabalho ambiciona contribuir para o avanço da SAT, através da exploração e integração de novas abordagens e técnicas de PLN, como o LDA [8] ou os *word embeddings* [9], não omitindo as abordagens mais convencionais, baseadas em grafos ou só na frequência e informação dos termos. Podemos ainda pensar em redução de frases e inclusão de heurísticas de transformação de texto, por exemplo ao nível sintático. Em suma, existe uma grande variedade de abordagens, técnicas e recursos que se podem combinar para gerar um sumário. A grande questão é saber que combinações destas é que tendem a gerar melhores resultados. Existem muitos trabalhos, mas a maioria foca-se exclusivamente em uma ou duas abordagens. Importa explorar uma combinação mais vasta e potencialmente mais benéfica.

## Objetivo

Atendendo ao vasto leque de possibilidades e abordagens existentes para a realização de SAT, importa explorar as combinações que produzem os melhores resultados, quer em termos quantitativos, quer qualitativos. Este é o principal objetivo deste trabalho, que deverá depois ser devidamente comunicado à comunidade científica. A avaliação será feita através das métricas (e.g. ROUGE) e conjuntos de dados (e.g. DUC, TAC) existentes, permitindo assim a obtenção de resultados objetivos e comparáveis com outros trabalhos.

## Tarefas Principais

Este trabalho contempla as seguintes tarefas principais:

**Tarefa 1:** Explorar o estado da arte no domínio da SAT, escrevendo a correspondente secção na dissertação.

**Tarefa 2:** Reunir e experimentar as diversas abordagens, técnicas e recursos, identificados na Tarefa 1, incluindo os dados e métricas de avaliação de sistemas SAT.

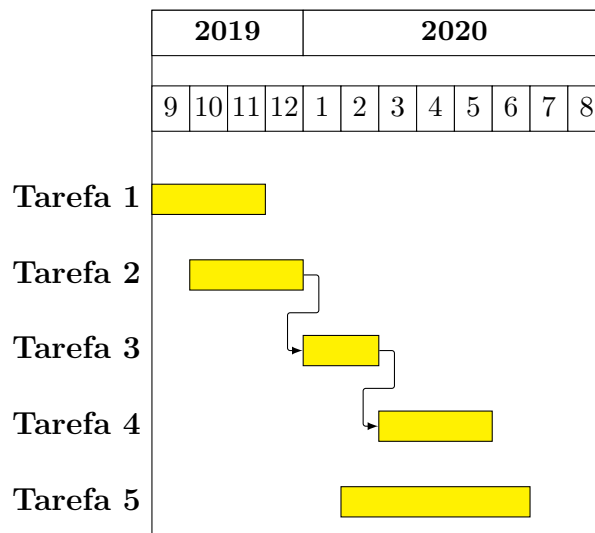
**Tarefa 3:** Explorar combinações de abordagens SAT, começando com um sistema base simples e progredindo para as combinações mais complexas e técnicas mais refinadas.

**Tarefa 4:** Pensar e implementar a integração de heurísticas de simplificação de frases, medindo os resultados obtidos.

**Tarefa 5:** Escrita da dissertação.

## Calendarização

Em termos de calendarização das tarefas principais, o trabalho está pensado para ter a seguinte distribuição temporal:



## Contactos

João Paulo Cordeiro (Gabinete 4.3)  
UBI, Departamento de Informática  
Rua Marquês d'Ávila e Bolama  
6201-001 Covilhã

## References

- [1] M. Gambhir and V. Gupta, “Recent automatic text summarization techniques: a survey,” *Artificial Intelligence Review*, vol. 47, no. 1, pp. 1–66, 2017.
- [2] O. Buyukkokten, H. Garcia-Molina, and A. Paepcke, “Seeing the whole in parts: text summarization for web browsing on handheld devices,” in *10th International WWW Conference*, 2001.
- [3] J. Liu, S. Seneff, and V. Zue, “Harvesting and summarizing user-generated content for advanced speech-based hci,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 6, no. 8, pp. 982–992, 2012.
- [4] H. P. Luhn, “The automatic creation of literature abstracts,” *IBM Journal of research and development*, vol. 2, no. 2, pp. 159–165, 1958.
- [5] H. P. Edmundson, “New methods in automatic extracting,” *Journal of the ACM (JACM)*, vol. 16, no. 2, pp. 264–285, 1969.
- [6] I. Mani, *Advances in automatic text summarization*. MIT press, 1999.
- [7] C.-Y. Lin, “Rouge: A package for automatic evaluation of summaries,” in *Text summarization branches out*, pp. 74–81, 2004.
- [8] D. M. Blei, “Probabilistic topic models,” *Communications of the ACM*, vol. 55, no. 4, pp. 77–84, 2012.
- [9] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” in *Advances in neural information processing systems*, pp. 3111–3119, 2013.