Universidade do Minho
Escola de Engenharia
Departamento de Informática

Bruno Vilas Boas da Silva

Extração Automática de Ontologias em Textos de Culinária não Estruturados

Outubro 2022

Universidade do Minho
Escola de Engenharia
Departamento de Informática

Bruno Vilas Boas da Silva

Extração Automática de Ontologias em Textos de Culinária não Estruturados

Dissertação de Mestrado
Mestrado Integrado em Engenharia Informática

Trabalho realizado sob orientação de
Professor Doutor Orlando Manuel de Oliveira Belo
Professora Doutora Anabela Leal de Barros

Outubro 2022

DIREITOS DE AUTOR E CONDIÇÕES DE UTILIZAÇÃO DO TRABALHO POR TERCEIROS

DECLARAÇÃO DE INTEGRIDADE

Declaro ter atuado com integridade na elaboração do presente trabalho académico e confirmo que não recorri à prática de plágio nem a qualquer forma de utilização indevida ou falsificação de informações ou resultados em nenhuma das etapas conducente à sua elaboração.

Mais declaro que conheço e que respeitei o Código de Conduta Ética da Universidade do Minho.

AGRADECIMENTOS

Em primeira instância, deixo o meu total agradecimento ao professor Doutor Orlando Manuel de Oliveira Belo pela supervisão do meu trabalho, pelo voto de confiança depositado em mim na realização desta dissertação e pela honestidade e franqueza que sempre transpareceu nas nossas conversas. De igual modo, à professora Doutora Anabela Leal de Barros por me permitir desenvolver este trabalho sobre um caso de estudo de sua autoria e pelas sugestões dadas para melhorar o sistema desenvolvido.

A nível pessoal, gostaria de agradecer aos meus pais por me presentearem com a oportunidade de evoluir, não só como pessoa, mas também como profissional e por todo o apoio prestado durante todo o percurso académico. Agradeço ainda ao meu irmão e restante família que me apoiaram durante todos estes anos.

Deixo ainda um agradecimento especial à minha namorada, que sempre me motivou e incentivou durante todo o processo de evolução no ensino superior.

E, por fim, um sincero obrigado à Universidade do Minho e a todos aqueles que me ajudaram, direta ou indiretamente, a alcançar os meus objetivos.

RESUMO

A resolução de problemas no âmbito de um domínio específico pode adotar técnicas e ideologias distintas. Para tal, é vital e imperativo elaborar uma análise contextual a todos os elementos pertencentes à teia de relações entre conceitos. Nesse sentido, o uso de uma ontologia permite construir uma rede semântica, no qual a mais importante premissa é a correta identificação dos conceitos e respetivos atributos. A automatização do processo de extração de ontologias permite construir ontologias mais escaláveis e uniformes, extraindo conhecimento assente nas mesmas premissas e padrões. No plano geral, uma extração automática facilita a análise e a leitura de informação de um problema apresentado numa linguagem própria. O trabalho desta dissertação focou-se na extração de conhecimento em textos não estruturados, mais concretamente, textos de culinária, com o intuito de disponibilizar uma ontologia que espelhasse o conhecimento interligado entre receitas. O verdadeiro desafio passa pela correta identificação de termos relevantes, com base em análise sintática, semântica, e linguística em geral, e pela formalização de relações entre os mesmos. A utilização de mecanismos de controlo e de automatização permitiu a extração do conhecimento presente nos textos não estruturados. Estes mecanismos foram aplicados conforme as características linguísticas inerentes aos documentos e restrições de domínio. A ontologia gerada pode ser consultada através de uma plataforma web, na qual o utilizador pode pesquisar os documentos importados no sistema e analisar a interligação entre receitas através da pesquisa por termos e por hiperligações que se encontram nos detalhes de cada registo de culinária.

PALAVRAS-CHAVE Ontologias, Textos Não Estruturados, Processamento de Linguagem Natural, Text Mining, Extração Automática, Análise de Textos.

ABSTRACT

The resolution of problems within a specific domain may adopt distinct approaches. As such, it is vital and imperative to elaborate a context analysis to each and every single existing element in the domain. For that matter, the use of an ontology allows the construction of a semantic environment where the most important factor is the correct identification of the concepts and its attributes. The automation of the whole process enables the ability to create more scalable and sustainable ontologies while extracting knowledge based on the same premises and patterns. An automatic extraction eases the analysis and understading of the information presented in a problem, usually written in natural language. This dissertation takes focus on the knowledge extraction in unstructured texts — culinary texts to be precised — with the sole goal of generating an ontology that exposes the knowledge intertwined between recipes. The main challenge presents itself as identifying the correct relevant terms, based upon context analysis and linguistics, and formalizing the relations among them. Using the proper control and automation mechanisms ensure the best results when retrieving knowledge from unstructured texts. Those mechanisms are chosen regarding linguistic characteristics and the corpus domain.

The generated ontology will be used as the backend of a web platform, where the user may search for the desired recipes imported in the system. Thus, the connection between recipes is highlighted when searching for a specific term and the hyperlinks embedded in recipe detailed information.

KEYWORDS Ontology, Unstructured texts, Natural Language Processing, Text Mining, Domain, Automatic Extraction, Text Analysis.

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE TABELAS

LISTA DE SIGLAS E ACRÓNIMOS

CSS	*Cascading Style Sheets*
BD	Base de Dados
UI	*User Interface*
NLP	*Natural Language Processing*
MVC	*Model-View-Controller*
OWL	*Ontology Web Language*
RDF	*Resource Description Framework*
POS	*Part-of-Speech*
ACID	Atomicidade, Consistência, Isolamento e Durabilidade
URI	*Universal Resources Identifier*
SPARQL	*SPARQL Protocol and RDF Query Language*
ILP	*Inductive Logic Programming*
FOL	*First Order Logic*

1

INTRODUÇÃO

1.1 CONTEXTUALIZAÇÃO

Desde o aparecimento do conceito de web semântica, os modelos ontológicos foram catapultados para a ribalta e passaram a ser amplamente usados em diversos sistemas (Davies & van Harmelen, 2003). Desde então, têm sido realizados vários estudos na área das ontologias, pelo papel determinante que estas desempenham na organização de dados e na representação da relação que estes estabelecem entre si, em qualquer área ou domínio de aplicação. Este progresso criou a necessidade de automatizar a criação do modelo de dados semânticos, geralmente representados em grafos, para substituir o moroso processo manual de construção ontológica. Explicado o surgimento de sistemas de extração automática de ontologias, resta caracterizar e descrever a evoluçdiacrónica do processo de extração de conhecimento, como expectativas de uma melhor qualidade dos dados coletados.

A mineração de dados é uma área em constante desenvolvimento, emergindo e desenvolvendo-se assente em novas técnicas e diretrizes, que permitem não só evidenciar o que é informativo, mas também atribuir um conjunto de propriedades individuais a cada pedaço de informação. Outrora, era suficiente e satisfatório extrair e processar componentes textuais de uma forma meramente linear, contrariando soluções e arquiteturas atuais que beneficiam largamente da correlação de dados.

As receitas surgiram como um meio de armazenar um conjunto de elementos e de processos, com o simples objetivo de padronizar e simplificar um conjunto de ideias do mesmo domínio aplicacional. O que inicialmente se registava em manuscritos foi informatizado e armazenado digitalmente. Isto potenciou, não só a capacidade de guardar mais receitas, como também a possibilidade de as consultar em qualquer lado. Para aperfeiçoar esta última etapa, nasceu a ideia e a necessidade de integrar os dados num sistema que os conseguisse relacionar, permitindo, assim, poupar horas de pesquisa em documentos meramente armazenados, sem qualquer tipo de correlação entre eles.

Uma solução para cumprir este requisito é a criação de um sistema de extração automática de ontologias, que seja responsável pela persistência da informação e a semântica que lhe é inerente.

Uma ontologia é um artefacto computacional para representação de conhecimento através de relações e classes. Estas bases de conhecimento requerem demasiado esforço humano para serem construídas devido à necessidade de peritos do domínio e de engenheiros de conhecimento. Um sistema de extração automática define-se como um conjunto de regras e axiomas que, aplicados à matéria-prima, resultam numa organização sistemática da informação, suscetível de ter um crescimento exponencial. Este conjunto de regras e axiomas será definido conforme as propriedades da área aplicacional. Um sistema de extração automática de ontologias procura definir um modelo de dados, construído através dos termos obtidos e processados pelo sistema de extração, que se caracteriza como um conjunto de conceitos interligados de um determinado domínio de aplicação. Esta extração ocorre de uma fonte de conhecimento que pode surgir sob a forma de conteúdos multimédia, páginas web, bases de dados, ou, como se verifica no âmbito desta dissertação, textos não estruturados. Em suma, estes sistemas são desenvolvidos sobre qualquer tipo de fonte de dados, o que evidencia um vasto leque de possíveis abordagens a considerar no processo de extração.

Nesta dissertação, é utilizada uma metodologia que permite construir um sistema escalável, sistemático e evolutivo. Este resultado é obtido através de uma análise efetuada em palavras e frases, associadas a conceitos, que formam a terminologia de um domínio. O reconhecimento da terminologia do domínio vincula-se ao conhecimento da linguística e de técnicas de extração, de entre os quais a identificação sintática (POS) é preponderante. O resultado final apresenta-se como um sistema capaz de retirar grande parte do esforço humano do processo de extração, de oferecer mecanismos de processamento que apliquem as mesmas regras a cada extração e suscetível de produzir resultados gradualmente melhores conforme o número de extrações já efetuadas.

1.2 MOTIVAÇÃO E OBJETIVOS

A extração automática de ontologias tem como objetivo a representação das relações semânticas entre os vários componentes presentes num dado conjunto de dados. Hoje, é vulgar optar pela realização desse trabalho em textos não estruturados, dada a enorme riqueza do seu conteúdo. Neste domínio, um dos grandes propósitos das ontologias é identificar e caracterizar elementos relevantes nos conteúdos dos textos, bem como estabelecer as relações entre eles. Depois, através de mecanismos de controlo e automatização, garantir a execução de processos de análise rápidos e eficazes dos conteúdos abordados e tratados nesses textos. Numa primeira fase deste projeto de mestrado, pretendia-se efetuar uma análise de um conjunto vasto de receitas de culinária antigas, disponíveis no livro "As Receitas de Cozinha de um Frade Português do Século XVI" (Barros, 2013), que apresenta o segundo manuscrito de culinária mais antigo que se conhece em Portugal (o primeiro caderno do códice 142 do Arquivo Distrital de Braga), inédito e desconhecido até essa data, com cerca de trezentas receitas, cuja edição semidiplomática e interpretativa, com estudo introdutório e glossário, é da autoria de Anabela Leal de Barros, com vista à conceção e desenvolvimento de um sistema de extração de conhecimento que nos permitisse obter, de forma (semi)automática, uma ontologia concreta sobre a culinária portuguesa. Com este sistema de análise de receitas, torna-se possível, então, relacionar os vários ingredientes ou processos culinários das variadas receitas, com a finalidade de serem estabelecidos padrões culinários do passado. A ontologia idealizada, bem como os seus respetivos mecanismos de exploração, deveria reconhecer esses padrões culinários, através de elementos como classes, propriedades, tipos e entidades chave dos registos analisados.

Após a análise e identificação dos termos culinários, sejam estes ingredientes, utensílios, processos culinários ou apontamentos gastronómicos, ter-se-ia que conceber e implementar um dicionário específico para a sua classificação e caracterização. Posteriormente, com vista à exploração dos conteúdos dos textos analisados, dever-se-ia desenvolver um angariador (*wrapper*) de textos específico, que permitisse extrair a informação contida nos textos de culinária, procedesse à sua análise e limpeza e, por fim, a armazenasse na base de dados orientada por grafos do sistema. A conceção (semi)automática da ontologia pretendida seria sustentada por um sistema de mineração de textos não estruturados que caracterizasse os elementos mais relevantes (classes, propriedades, tipos e entidades). Neste momento, podemos afirmar que o sistema ontológico que foi desenvolvido, assente nos objetivos e pressupostos enunciados, possibilita o acesso à caracterização da informação contida nos textos das receitas culinárias a que tivemos acesso ao longo do desenvolvimento dos trabalhos desta dissertação.

1.3 TRABALHO REALIZADO

A fim de perceber a problemática em que assenta esta dissertação, numa fase inicial, foi efetuada uma pesquisa detalhada dos sistemas de extração automática de ontologias. O objetivo desta investigação foi perceber quais as características inerentes a estes sistemas, assim como conhecer os conceitos e processos que se desenrolam durante o seu desenvolvimento. Após a análise do estado da arte, a investigação passou a ser cada vez mais específica e voltada para cada etapa no processo de extração de conteúdos, mais concretamente, em textos não estruturados. Nesta etapa, exploraram-se as vantagens e desvantagens que determinados processos (e.g. tradução da linguagem natural presente nos textos, abstração de género e número dos termos extraídos, uso da frequência relativa dos termos extraídos, persistência dos dados na ontologia, etc...) trariam ao sistema, com o intuito de perceber quais se enquadrariam melhor no sistema posteriormente desenvolvido. Uma vez estudados todos os cenários (abordagens estatística, top-down e bottom-up), iniciou-se a fase de análise do caso de estudo.

Na análise deste caso de estudo — "As Receitas de Cozinha de um Frade Português do Século XVI"—, foi importante identificar, em primeira instância, o conjunto das características mais relevantes do domínio em causa: a culinária. Estas características são reveladoras de propriedades quase transversais a todos os documentos deste domínio, nomeadamente ao corpo de uma receita. Após este breve levantamento de propriedades, procedeu-se ao estudo de vários registos de culinária presentes no texto editado do manuscrito alvo desta dissertação, tendo-se optado pela lição interpretativa (com grafia atualizada), já que esta permite ultrapassar algumas das dificuldades relacionadas com a ampla variação ortográfica que podem apresentar alguns manuscritos em português clássico. A intenção era simples: perceber o que o sistema de extração automática poderia explorar e idealizar a ontologia que resultaria desse processo.

Concluída a fase de estudo e análise do problema, concebeu-se uma plataforma que pudesse ser alimentada por um sistema de extração automática de ontologias. Esta plataforma, desenhada como uma aplicação *Web*, permite, a cada utilizador, importar receitas do manuscrito estudado e, posteriormente, pesquisar conteúdos utilizando a ontologia. Terminada a preparação da plataforma para acolher este sistema, começou a fase de desenvolvimento do processo de extração de conteúdos. Esta etapa foi marcada pelo desenvolvimento e, simultaneamente, teste de diferentes abordagens da extração em textos não estruturados, a fim de perceber qual a metodologia que produziria melhores resultados. No final do desenvolvimento do sistema de extração e geração da ontologia, trabalhou-se nas funcionalidades da plataforma e no seu ambiente gráfico.

1.4 ESTRUTURA DA DISSERTAÇÃO

Para além do presente capítulo, esta dissertação integra mais quatro capítulos, nomeadamente:

- Capítulo 2 — Extração Automática de Ontologias, em que se apresentam definições concretas de conceitos abordados ao longo da dissertação, nomeadamente as propriedades de uma ontologia, a evolução do processo de extração em textos e, também, algumas ferramentas e metodologias adotadas nesta temática. Além disso, evidenciam-se, também, os benefícios retirados deste tipo de implementação, na qual se formaliza uma ontologia com base nas regras e axiomas do domínio e se caracteriza a mesma através da informação extraída dos textos não estruturados, assim como potenciais vulnerabilidades intrínsecas à extração automática de ontologias.
- Capítulo 3 — O Caso de Estudo, no qual se expõe detalhadamente o processo de desenvolvimento de um sistema de extração automática de ontologias em textos não estruturados, com aplicação concreta ao livro de cozinha incluído no manuscrito 142 do Arquivo Distrital de Braga, inédito e sem título nem autor (por falta do primeiro fólio), publicado sob o título *"As Receitas de Cozinha de um Frade Português do Século XVI"* (Barros, 2013). É ainda caracterizado todo o processo de tradução da linguagem natural para as estruturas de dados do sistema, assim como as regras e padrões que agem como mecanismos de controlo na extração do conhecimento. Por último, é esclarecida a forma como a informação é retida no sistema.
- Capítulo 4 — Análise de Resultados, capítulo no qual se apresenta a informação resultante do processo definido no capítulo anterior. Na prática, são evidenciadas as vantagens do uso do sistema ontológico construído, bem como se apresenta o ambiente de testes que foi utilizado, abordando os seus principais aspetos, desde a tecnologia utilizada no seu desenvolvimento até às funcionalidades que estão disponíveis. Além disso, é explicado o funcionamento da plataforma desenvolvida para divulgação dos resultados, desde a importação de receitas até à pesquisa de resultados na base ontológica criada.
- Capítulo 5 — Conclusões e Trabalho Futuro, o último capítulo, apresenta uma retrospetiva de tudo aquilo que foi feito, desde as decisões tomadas, em conformidade com os resultados obtidos em diferentes abordagens, até às melhorias que trariam mais fiabilidade à informação extraída. Além das conclusões, são referidos alguns aspetos a aperfeiçoar ou desenvolver no sistema, em trabalho futuro.

2

EXTRAÇÃO AUTOMÁTICA DE ONTOLOGIAS

2.1 ONTOLOGIAS

Uma ontologia, na sua essência, define um conjunto de conceitos relacionados entre si, com a finalidade de criar uma interoperabilidade entre eles e formar uma ligação semântica perante o domínio em causa. A semântica ilustrada pela ontologia é constituída por classes, propriedades e relações, sendo estas últimas que ligam os conceitos entre si. Depois disso, surgem as entidades, que são unidades representativas das classes, ou, em termos gerais, conceitos, com a finalidade de povoar a ontologia. Uma ontologia de domínio representa o conhecimento extraído de determinada linguagem, própria de um domínio específico, partindo da premissa de que cada conceito tem apenas um significado válido. Existem ainda ontologias superiores, que são definidas como um modelo de objetos comuns, que podem ser aplicados a diversas ontologias de domínio, ou seja, um conceito, numa ontologia superior, poderá representar vários conceitos em ontologias de domínio diferentes.

Relativamente às ontologias de domínio, as regras e axiomas que as compõem são determinantes para a definição dos conceitos e entidades presentes na ontologia. As regras são declarações que descrevem inferências lógicas, resultantes de afirmações impostas pela fonte de conhecimento. Axiomas definem-se como declarações (incluindo regras) representadas como formas lógicas que comprovam a ideologia que a ontologia representa no seu domínio aplicacional.

No âmbito da Web Semântica, surgiu a necessidade de criar uma linguagem que conseguisse verificar a consistência do conhecimento, assim como transformar conhecimento implícito em explícito. Essa linguagem foi a OWL, *Web Ontology Language* (Mankovskii *et al*., 2009). Esta linguagem permite uma distinção unívoca, e também detalhada, das classes, propriedades e relações. Além disso, com a OWL, a definição das classes e relações, juntamente com a respetiva ordenação hierárquica, resulta num enriquecimento da modelação ontológica em grafos semânticos, conhecidos como *RDF triplestore* (www.ontotext.com). Este define-se como uma base de dados em grafo que armazena toda a informação como uma rede de objetos e usa inferência para descobrir nova informação derivada de novas relações. As *triplestores* expressam uma relação entre sujeito, predicado e objeto. Este formato permite ligar um sujeito a qualquer objeto que cumpra a premissa estabelecida pelo predicado. O predicado pode surgir quer como uma relação quer como uma propriedade do sujeito. Teoricamente, a procura por uma receita, numa ontologia, que tenha um ingrediente específico, pode ser representada pela seguinte declaração:

- Sujeito — "Receita"
- Predicado — "tem"
- Objeto — "ingrediente"

O sujeito e o objeto apresentam-se como dois nodos, o que significa que o predicado é representado pelas ligações que os unem. Um dos aspetos negativos deste formato é a impossibilidade de acrescentar descrições ou declarações nas arestas. No entanto, esta limitação já não se verifica em versões mais atuais. O uso de declarações sobre relações auxilia na compreensão da ligação entre duas entidades na ontologia.

Para ter acesso à *triplestore* desejada, é necessário utilizar um identificador que caracteriza a mesma de forma distinta de todo o universo de triplos disponíveis na Web. O conceito nuclear do formato *RDF triplestore*, tal como no paradigma de semântica de dados, denomina-se *Universal Resources Identifier* (URI). Tal como existem chaves primárias para se identificarem registos, univocamente, numa tabela de um modelo relacional, o URI foi criado com o propósito de servir como um identificador único e universal na Web. Este identificador é que permitirá, não só, usar o triplestore desejado, mas também executar queries sobre a informação interligada e contida no grafo semântico.

Para efetuar consultas em qualquer base de dados que consiga ser mapeada para RDF, utiliza-se *SPARQL Protocol and RDF Query Language* (SPARQL). Esta linguagem padronizada, especialmente para RDF, integra uma funcionalidade de consulta de padrões semânticos juntamente com as suas conjunções e disjunções. O resultado das *queries* executadas surgirão como um conjunto de registos ou como grafos RDF.

2.1.1 *Utilidade de uma Ontologia*

A aplicabilidade de uma ontologia é imensa, sendo transversal a qualquer problema existente no contexto do mundo real. As áreas de aplicação podem ir desde as ciências da saúde ao comércio eletrónico, ou qualquer outra área que necessite de explorar informação partilhável e reutilizável dentro do seu domínio. Além de ser uma forma escalável e sustentável de guardar informação, uma ontologia tem como objetivo uma melhor gestão do conhecimento extraído de um determinado domínio. Uma ontologia é capaz de relacionar os atributos, e relações, inerentes a um conceito e vice-versa, permitindo uma análise e compreensão mais rápidas do que é exposto na fonte de conhecimento.

O maior benefício que se pode obter na utilização de uma ontologia decorre do facto de os conceitos estarem interligados, funcionando como um mecanismo automático de raciocínio. À semelhança do cérebro humano, uma ontologia trabalha e raciocina sobre todo o enredo de conceitos, criando uma perceção real daquilo que o domínio apresenta. Para lá desta grande vantagem, segue-se a versatilidade com que se pode adaptar uma ideologia a uma ontologia; tal é o caso das variadas áreas de aplicação que este sistema integra. Se, por um lado, uma ontologia pode representar uma variedade de domínios diferentes, por outro lado, pode apresentar-se como uma desvantagem o facto de, aquando da construção da mesma, as restrições impostas pelas regras e axiomas poderem limitar a integração de novos dados no sistema. Quanto a este último ponto, o ideal será estudar a fonte de conhecimento e a sua estrutura, para que a formalização da solução seja transparente e respeitadora das premissas que são impostas pelo domínio.

Relativamente à temática a ser abordada pela ontologia, o uso desta última permite uma distinção denotativa de todos os dados integrados no universo semântico. Idealmente, cada palavra deveria exprimir um significado único, mas tal não acontece, devido a um simples fator: o contexto. As ontologias assumem o contexto traduzido pelo domínio e atribuem o significado correto às palavras ou terminologia. Esta característica oferece uma melhoria na qualidade da classificação dos conceitos a definir, o que, na prática, se traduz numa redução da ambiguidade. A definição de regras e axiomas, no âmbito de um domínio em particular, só deverá ser efetuada uma vez, o que possibilita a inserção de conhecimento por parte de outros agentes da mesma área de domínio. Por exemplo, no caso da área da biomédica, a partilha de informação através da mesma ontologia potencia a inferência de novos resultados e de novas descobertas, ou na culinária, em que a descoberta de padrões gastronómicos traduz um vasto leque de costumes de uma região, além de enriquecer, e possivelmente aperfeiçoar, receitas existentes.

2.1.2 *Exemplo: Ontologia de Receitas*

Atualmente existem milhares de receitas culinárias espalhadas pelo globo, sendo muitas delas parte integrante da cultura de uma região, tornando-se mesmo tradição em algumas épocas do ano. Cada receita tem vários elementos importantes, e a sua combinação torna cada resultado num prato de gastronomia diferente de todos os outros. Para confecionar tais pratos, é necessário que a receita exponha os ingredientes e quantidades dos mesmos, bem como os utensílios do âmbito da sua confeção. Esta parte é vital na preparação da receita, bem como na garantia de que todo o procedimento possa ser efetuado. A seguir, será necessário saber como é que estes ingredientes vão ser conjugados e tratados, sendo esta parte representada por processos. Além de tudo isto, cada processo está associado a um determinado espaço temporal, assim como a totalidade da receita. Imaginemos uma ontologia que descreva uma receita. Sem dúvida que *Ingrediente* seria o conceito mais óbvio nesta ontologia, mas, além deste, é necessário investigar que outros fatores se relacionam com este conceito e desempenham papel importante no âmbito da culinária. Para além dos ingredientes, identificam-se como conceitos relevantes: Tempo, Quantidade, Preparação, Utensílios, Processos, Etapa e Receita. Depois de identificados todos os termos relevantes, será necessário estabelecer propriedades do objeto, isto é, como é que o conceito se relaciona com os outros, e atributos do objeto, que por norma traduzem-se em características do conceito. Por fim, iniciar-se-á o processo de construção da teia semântica que irá dar origem a uma ontologia que descreve o domínio de Receitas, assim como o papel que cada entidade participante representa e desempenha. O objetivo principal é perceber o fluxo de dados que cada classe, vulgo entidade, providencia à ontologia e o impacto que terá caso a mesma não esteja bem definida.

2.1.3 *Utilidade do Exemplo*

O caso descrito apresenta algum do raciocínio necessário para dar origem a uma ontologia, normalmente utilizado por engenheiros do conhecimento ou especialistas em ontologias. Este é o papel que um sistema de extração automática de ontologias pretende substituir, não só para poupar algum tempo na criação de ontologias, mas também para mitigar o erro humano que está presente na criação manual de uma ontologia. Para além do lado mais técnico, a ontologia de receitas exemplificada permite elucidar sobre os conceitos inerentes ao domínio da culinária e como os mesmos se relacionam. Mais ainda, a ontologia apresentada serve como um guião para o que deverá ser a ontologia a ser gerada pelo sistema de extração automática desenvolvido ao longo desta dissertação. Este exemplo tem como finalidade mostrar a facilidade com que a informação é exposta, permitindo uma melhor pesquisa de receitas. Na prática, poderá servir a uma pessoa que tenha uma

vaga ideia do procedimento da confeção de uma receita, mas que, com outra receita que partilhe os mesmos ingredientes, já tenha reunidas todas as condições para confecionar o prato. Aplicando a mesma lógica aos processos, quem está encarregado de confecionar poderá querer um prato que tenha obrigatoriamente que possuir um elemento frito, ou grelhado. Aplicando isto a todas as classes da ontologia, permitirá ao utilizador poupar muito tempo de pesquisa, pois a semanticidade e congruência do sistema ontológico assim o permite. Uniformizando cada receita, a homogeneidade será transversal a toda a base de conhecimento integrada no domínio da culinária.

2.2 SEMÂNTICA E INTEROPERABILIDADE

As várias conotações que uma palavra adquire estão sempre dependentes da linguística e estão diretamente relacionadas com a forma como o autor da frase as utiliza e com o contexto em que as insere. Este aspeto pertence ao âmbito da semântica. A interpretação de palavras, frases ou sinais determina a nossa compreensão do discurso oral e escrito, em geral, a nossa compreensão dos outros. O resultado desta compreensão manifesta-se nas decisões que são tomadas e influenciadas pela conotação que cada pessoa atribui a cada palavra. A semântica pode ainda ser afetada na linha temporal, e espacial, conforme se vão aprendendo ou criando novas conotações e perdendo outras; trata-se da evolução semântica a que estão sujeitos os lexemas.

A interoperabilidade define-se pela habilidade que diferentes sistemas, aplicações ou dispositivos têm para se conectarem e comunicarem entre si sem que haja um esforço em particular para que tal aconteça. Este conceito não só permite a existência de projetos em larga escala como expande horizontes, e é a razão pela qual existem sistemas modulares num mundo cada vez mais diversificado de soluções direcionadas para propósitos específicos, mas que são capazes de trocar informação com sistemas completamente distintos (Meslati *et al*, 2019). No fundo, a interoperabilidade permite a livre transmissão e acesso de dados presentes em diversos sistemas, desconsiderando o autor ou a origem dos dados.

É importante salientar que existem ainda vários tipos de interoperabilidade, dos quais são dignos de nota os seguintes (Mucheroni & Modesto, 2011):

- **Interoperabilidade sintática** — Caracteriza-se por ser uma funcionalidade que permite com que os sistemas consigam comunicar entre si através de protocolos e formatos compatíveis. Para facilitar esta tarefa, foram criados formatos universais, como é o caso do XML, JSON, CSV, *etc*. Este tipo de interoperabilidade pode também ser mencionado como interoperabilidade estrutural.

- **Interoperabilidade semântica** — Capacidade que permite aos sistemas trocar e, de forma precisa, interpretar informação de forma automática. Tal acontece quando a estrutura dos dados é uniforme entre todos os sistemas envolvidos.
- **Interoperabilidade entre domínios ou organizações** — Neste tipo de interoperabilidade são definidas políticas, boas práticas e requisitos entre sistemas distintos. Este caso em particular não prevê, necessariamente, uma estrutura de dados homogénea, nem protocolos de utilização obrigatória para o funcionamento e interoperabilidade entre plataformas. Em suma, são apenas fulcrais os aspetos que não sejam técnicos.

A interoperabilidade tecnológica, aliada à semântica, permite reutilizar recursos já existentes, além de combinar esforços para um objetivo final e comum: estabelecer uma qualidade transversal entre sistemas. Com a emergência da necessidade de compatibilizar sistemas relativamente à troca de informação, as ontologias passaram para o primeiro plano como uma solução para problemas de interpretação de dados. O mesmo acontece na informação que circula entre sistemas. Nesse sentido, as ontologias procuram agrupar conjuntos de conotações de um conceito para facilitar a compreensão, permitindo ao sistema herdar uma terminologia.

2.3 EXTRAÇÃO DE ONTOLOGIAS

2.3.1 *Extração Semiautomática de Ontologias*

A construção manual de uma ontologia é muitas vezes demasiado demorada para os peritos no domínio, pois requer uma larga análise do contexto em causa e a correspondente identificação de todos os elementos pertinentes para a constituição ontológica. A pesquisa desta temática, nos dias de hoje, recai maioritariamente sobre a construção semiautomática e automática, de maneira a ultrapassar os problemas identificados no processo de construção manual da ontologia. A extração (semi)automática de ontologias, ou *Ontology Learning* (Mishra & Jain, 2014), caracteriza-se pela recolha dos termos do domínio e relações entre conceitos que os termos representam, a partir de uma fonte de conhecimento de textos escritos em linguagem natural (Choudhary & Tomar, 2014). O resultado deste processo é convertido, posteriormente, numa linguagem ontológica que permitirá uma rápida pesquisa da informação catalogada.

A primeira etapa que caracteriza este processo passa pela extração da terminologia do domínio. É aqui que são extraídos e catalogados todos os termos que poderão, ou não, ser relevantes para a ontologia. Nesta lista de possibilidades, cabe a um perito do domínio distinguir o que é ou não relevante no âmbito da temática, ou então, atribuir esta responsabilidade a um conjunto de mecanismos de controlo e distinção, através de análise

estatística ou de padrões textuais. Ainda nesta fase, todos os sinónimos são agrupados pelo simples facto de partilharem o mesmo significado, consequentemente, são caracterizados pelo mesmo conceito.

De seguida, procede-se à caracterização dos conceitos (Figura 1), que não são mais do que as instâncias representativas dos termos extraídos. Os conceitos são estruturados de forma taxonómica, tipicamente recorrendo a regras parte-todo. Estas regras verificam se o termo é uma instância de um conceito previamente definido e são responsáveis pela validação de uma subclasse desse conceito. No entanto, raramente estas verificações serão triviais, pois a definição das mesmas envolve uma compreensão mais complexa. Além destas restrições, devem estar aliado um conjunto de propriedades, de preferência reveladoras de um conceito específico, que permita colmatar a, por vezes, deficiente caracterização de uma subclasse.

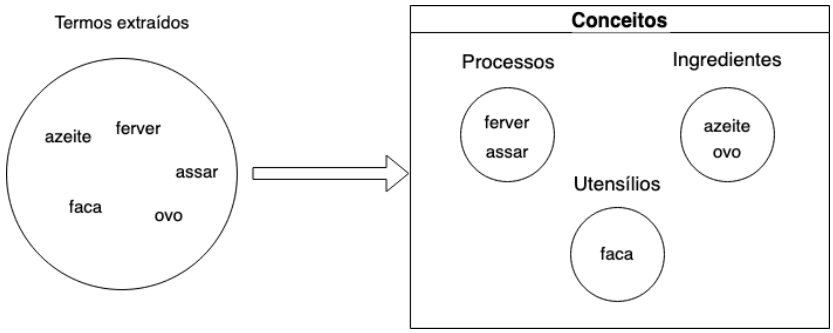

Figura 1: Classificação dos termos em conceitos

Uma vez que todos os conceitos sejam identificados, segue-se a identificação das relações não taxonómicas que estabelecem um elo de ligação entre conceitos. Esta tarefa pode ser feita através de associações, que são definidas com base em expressões ou padrões característicos da linguagem natural. Uma outra abordagem passa pela extração de verbos, que tendencialmente representam ações de um termo sobre outro, que permitem inferir uma relação entre conceitos. A precisão que ambas as abordagens apresentam carece de testes e avaliação de acordo com o caso de estudo.

A derradeira etapa assinala um conjunto de regras que caracterizam todo um conjunto de conceitos. Através de uma análise sintática da linguagem natural, determinam-se certas dependências e características que definem essas mesmas regras. O resultado do uso destas regras traduz-se numa lista de axiomas que descrevem formalmente os conceitos.

2.3.2 *Extração em Textos Semiestruturados*

Uma ontologia é tão boa quanto a qualidade da informação que a define, e em relação a esta última, a responsabilidade recai inteiramente sobre o processo de extração, e neste caso em concreto, de textos semi estruturados. Textos semiestruturados estão geralmente associados a informação escrita e apoiada em linguagem de marcas, vulgo *tags*, que facilitam a identificação dos conceitos que farão parte da ontologia. Além destas marcas explícitas no texto, o documento a ser analisado pode ainda incluir uma composição estrutural homogénea que possa evidenciar um significado taxonómico. A ideologia aplicada nesta extração é simples – as marcas identificam univocamente o que o termo representa –, e evita, assim, recorrer a técnicas de extração mais elaboradas, que se apoiam na análise contextual. Neste caso, é obtido o conhecimento apresentado na fonte sem que o sistema saiba necessariamente todo o conteúdo e contexto textual. Uma vez coligidos todos os termos marcados, a sua relação com os restantes será quase transparente e a ontologia será formada em etapas sucintas.

Apesar da aparente panóplia de vantagens associadas a este tipo de textos, as desvantagens são proporcionalmente mensuráveis em esforço necessário para transformar o texto escrito em linguagem natural em texto semiestruturado. Outra desvantagem igualmente evidente reside no facto de ter que se redesenhar todo o processo de parsing, caso a linguagem natural sofra alterações ou passe a incluir novos elementos.

2.3.3 *Automatização da Extração de Conhecimento em Textos não Estruturados*

A extração de conteúdo pertencente ao domínio é, em grande parte das vezes, morosa, obrigando a quem detém a capacidade de referir e identificar objetos relevantes para a ontologia, a uma nova análise contextual sempre que surge uma nova fonte sobre o mesmo domínio. Como resposta a este problema, e também por forma a uniformizar o controlo e extração do conhecimento, iniciou-se a automatização deste processo. Ao contrário do que acontece em textos semiestruturados, os termos relevantes para a ontologia não estão marcados, o que torna a fase analítica do contexto um pouco mais complexa. Enquanto em textos semiestruturados a extração é direta, neste tipo de estrutura textual será necessário recorrer a técnicas de processamento de linguagem natural e de text mining. O objetivo será sempre aproximar dos 100% a eficácia de identificação e extração dos conceitos predominantes, e para isso será necessário analisar e tratar cuidadosamente a informação apresentada e não estruturada. É importante salientar que a ontologia só fará sentido se as palavras-chave forem todas identificadas, caso contrário, a fase seguinte acontecerá de forma defeituosa. Posteriormente, uma vez que os conceitos estejam analisados e corretamente

identificados, são então criadas as relações que darão valor à ontologia e permitirão uma rápida abordagem e conhecimento de todo o conteúdo presente no domínio.

No que toca aos processos de extração em si, podem ser adotados vários métodos, desde a caracterização de padrões textuais e a extração de conteúdos com base na frequência relativa das palavras até à extração de elementos linguísticos próprios do domínio ou do autor, sendo todos estes processos relevantes para uma melhor qualidade no parsing e para a compreensão da área de conhecimento a ser estudada.

Em qualquer registo contemporâneo de culinária, o autor costuma dividir cada receita em vários blocos descritivos: o título da receita, os ingredientes e respetivas dosagens, e, por último, o procedimento. Além destes, poderão eventualmente surgir notas ou apontamentos que facilitem a leitura do público alvo. Por isso, é importante dividir a estrutura textual em vários blocos de informação para facilitar a definição do tradutor de linguagem natural em conhecimento explícito. Este é o caso do texto editado do códice de receitas do Arquivo Distrital de Braga, contudo, nas receitas antigas, e tal como elas surgem neste manuscrito, não se apresentam previamente todos os ingredientes e respetivas dosagens, nem outras notas explicativas ou esclarecedoras, o que foi acrescentado pelo editor e explicado nos critérios de edição (Barros, 2013). As receitas de séculos pretéritos eram frequentemente escritas num registo rápido, num único bloco, explicando mais os menos sucintamente (por vezes demasiado sumariamente) o procedimento, ao mesmo tempo que vão sendo nomeados os ingredientes, e nem sempre indicando quantidades e tempos, ou fazendo-o de modo pouco homogéneo e sistemático, já que as receitas, tendo valor patrimonial, eram colecionadas de várias fontes ao longo dos séculos. Na presente dissertação, utilizámos a versão interpretativa (com grafia atualizada) do texto já editado, com todos os dados organizados e inseridos manualmente pelo editor, nomeadamente os relativos aos ingredientes e quantidades, extraídos do bloco de texto de cada receita e colocados logo após o título da mesma, em tamanho de letra menor, para mais fácil orientação do leitor/utilizador.

Os algoritmos e dicionários a aplicar ao texto dependem fortemente da língua e da linguística, ou seja, do discurso usado, o que requer que haja uma interpretação da linguagem natural existente frase a frase, parágrafo a parágrafo. Para uma análise ainda mais detalhada, recorre-se à divisão do fluxo de carateres, isto é, guardam-se todos os elementos textuais em forma de tokens. Estas anotações poderão ser palavras, números ou até mesmo pontuação, caso seja relevante para o caso em estudo. A língua portuguesa dá uso a prefixos e sufixos para formar derivações de uma palavra primitiva, ou raiz, e no contexto ontológico, cada conceito deve ser identificado univocamente, rejeitando termos que possam criar ambiguidades.

A compreensão do documento requer uma análise tanto global como mais detalhada, ou, mais concretamente, sintática. Existem esses dois tipos de análise a serem efetuados aos textos não estruturados, a nível macro e a nível micro. A primeira foca-se numa compre-

ensão mais geral do documento, desde a classificação e catalogação da informação até à determinação da similaridade entre os vários textos presentes no documento. Relativamente à segunda análise, serve para extrair conteúdos mais concretos, tais como a identificação de entidades, regras linguísticas explícitas no texto e uma avaliação sintática das frases. É neste momento que se escolhe a abordagem a ter para a identificação da classe e da função gramatical das palavras: top-down, bottom-up ou estatística. A primeira analisa sintaticamente a frase, caracterizando individualmente cada termo presente na frase pela sua função sintática. Esta abordagem poderá revelar problemas diante de uma frase estruturalmente diferente. A abordagem bottom-up simplesmente cria padrões textuais e extrai conteúdo conforme esses padrões, o que se traduz na manutenção destes mesmos padrões para construtores de frase novos (El Ghosh *et al*., 2017). A análise estatística não difere muito da anterior, pois faz uso de correspondência de padrões com uma base de dados de padrões que foram gerados através do treino de dados, o que acontece em algoritmos de *machine-learning* e *big data* (Ahmad & Gillam, 2005). Esta abordagem requer um treino alargado dos dados, o que em certos casos não será viável, juntamente com o problema exposto na análise bottom-up.

2.3.4 *Vantagens e Desvantagens*

A perceção de todos os fatores é inevitavelmente morosa e requer uma análise cuidada do domínio, e, ainda assim, é por vezes uma tarefa tanto mais árdua quanto maior for a complexidade do problema. Neste sentido, um sistema de extração automática retira essa responsabilidade da entidade que domina a linguagem do domínio, sendo apenas necessário evidenciar as características implícitas na linguagem natural para que o sistema possa evoluir a partir daí. Revela-se igualmente benéfico o uso desta abordagem, por exemplo, em textos não estruturados muito extensos, em que o leitor provavelmente acaba a análise textual com um discernimento diferente do que aquele que possuía no começo da leitura, eliminando assim fatores externos que interferem com a qualidade da extração do conhecimento. Aliada à facilidade de análise e extração de novos conceitos, a automatização de todo este processo permite construir uma ontologia de forma controlada e escalável. Será controlada no sentido em que cada nova entidade introduzida terá cumprido todos os requisitos impostos pelas regras de domínio. O sistema torna-se escalável pela inferência de novos dados sobre as entidades existentes, o que implica que cada novo elemento adicionado à ontologia não será uma mera redundância ou ambiguidade (Horrocks, 2013).

Definir que tipo de informação se pretende extrair nem sempre é uma ciência exata, pelo que os fatores envolventes mudam e, consequentemente, levam a uma alteração das regras aplicadas no sistema. Quanto mais genérica for a especificação das propriedades do domínio, mais pobre será a qualidade da informação inferida pela ontologia. Se, por um lado, a especificação é genérica e sofre alterações, o impacto será baixo mas o resultado

final continuará a ser pobre, por outro lado, se a especificação é bem construída mas sofre alterações, causará impacto mas será, eventualmente, bloqueadora, o que implicará uma refatorização de todas as regras aplicadas.

A desvantagem mais evidente de um processo de extração automatizado em textos não estruturados resume-se ao facto de que a extração nunca será totalmente precisa. Em grande parte dos casos, a atribuição de um conceito a um termo está altamente dependente da análise sintática e de padrões morfológicos. Possuir um sistema capaz de identificar toda a variação e que esteja totalmente preparado para conseguir extrair todos os fatores corretamente é quase uma tarefa incessante. No entanto, é possível construir um mecanismo de extração automática com resultados suficientemente aceitáveis para tornar viável este tipo de sistemas.

2.4 ABORDAGENS E TÉCNICAS DE EXTRAÇÃO

As abordagens desenvolvidas na área de extração automática de ontologias visam obter o melhor resultado na caracterização dos termos, conceitos, relações e axiomas (Mishra & Jain, 2014). O desenvolvimento ontológico impõe a necessidade de retirar a terminologia de domínio. Tal é alcançado com o uso destas abordagens, podendo cada uma surtir melhores resultados conforme as características do objeto de estudo. Segundo Shamsfard (2003), existem essencialmente quatro abordagens a seguir:

- **Estatística** — Trata-se de uma abordagem que assenta no tratamento do input através de uma análise estatística. Esta metodologia está amplamente enraizada em técnicas de *machine learning* e *data mining*. A cada palavra, ou conjunto de palavras, são efetuadas operações de cariz estatístico com o intuito de apurar a sua relevância no contexto. Em alguns casos, a frequência com que uma palavra surge está diretamente relacionada com a preponderância que a mesma terá no domínio. Esta abordagem pressupõe um treino de dados prévio para que os resultados sejam minimamente viáveis, ou seja, uma análise prévia ao corpus do mesmo domínio. A extração de elementos para a ontologia será substancialmente melhor, com base num melhor fator de acerto à medida que a ontologia cresce. Este último ponto pode revelar-se uma contradição em casos de estudo nos quais a área de domínio é muito específica, ao ponto de não haver possibilidade de implementar um treino de dados prévio devido à inexistência de casos de estudo idênticos ao escolhido.

- **Lógica** — Atua com base em princípios de programação baseados em Lógica, tais como a Programação Lógica Indutiva (ILP), a Lógica de primeira ordem (FOL), a aprendizagem de regras e a aprendizagem proposicional. Apesar de ser importante a sua utilização para tarefas complexas, como a extração de regras e axiomas, a

sua aplicabilidade na caracterização de conceitos é pobre, pela ausência de um fator determinante nesta temática: o contexto.

- **Correspondência de padrões** — Captura um conjunto de padrões, léxicos e sintáticos, geralmente sob a forma de expressões regulares. Hearst apresentou padrões léxico-sintáticos para extrair de textos relações de hiponímia e hiperonímia. Estes são especialmente úteis para classificar conceitos, e respetiva hierarquia, numa ontologia. Uma abordagem deste tipo, idealmente, requer um caso de estudo com uma marca representativa distinta ou, no caso dos textos não estruturados, uma estrutura e escrita uniforme e homogénea (Hassan *et al.*, 2018). A manutenção destes padrões torna-se complexa, e por vezes pouco eficaz, quando o autor muda e, consequentemente, o estilo de escrita presente nos textos. Cada pessoa expressa-se de forma diferente, assim como a sua forma de escrever revela um cunho pessoal, pelo que os padrões léxicos e sintáticos produzem resultados diferentes. Contudo, se os padrões forem manualmente definidos, esta metodologia não carecerá de um treino de dados e tornar-se-á ideal para casos de estudo isolados.

- **Linguística** — Assume uma dependência linguística e extrai conhecimento ontológico a partir de texto em linguagem natural. O processamento de linguagem natural desempenha um papel fulcral nesta abordagem, pois interpreta e transforma os textos em conhecimento relevante. Normalmente, é efetuado um pré-processamento dos dados com o objetivo de identificar e caracterizar sintática e semanticamente todos os elementos presentes numa frase. Tipicamente dependentes de NLP, os processos de extração e text mining, neste caso em concreto, visam analisar a relação entre verbos, a estrutura das frases e expressões linguísticas. A construção dos conceitos e relações presentes no sistema estará, implicitamente, vinculada à linguagem natural expressa nos textos, o que torna o uso desta abordagem quase imprescindível.

Qualquer abordagem não tem que ser exclusivamente a única a ser aplicada no desenvolvimento de um sistema de extração automática de ontologias. Na verdade, o uso singular das metodologias definidas raramente produz os resultados desejados, devido à complexidade que a informação retida em textos de linguagem natural comporta. A escolha da metodologia passa por avaliar a sua viabilidade, conforme o domínio e caso de estudo em causa, e se os resultados previstos correspondem às expectativas.

Desde meados dos anos cinquenta, surgiram diversas abordagens no âmbito da extração de termos, conceitos e relações a partir de textos. No entanto, foi no fim do século XX que houve uma mudança significativa nas abordagens previamente sugeridas, devido à disponibilidade de novas técnicas de processamento de linguagem natural. Naturalmente, nos dias de hoje ainda são usadas essas mesmas técnicas, entretanto aprimoradas, na geração de ontologias.

Reinberger *et al*. (2004) sugeriu um método para extração ontológica a partir de textos que consistia na criação de uma estrutura ontológica primária que posteriormente seria refinada por analistas. De acordo com o seu método, numa primeira instância, um documento do mesmo domínio é analisado por um analisador mais genérico. Este analisador separa todos os elementos presentes no texto, analisa a sua função sintática, determina o início e fim de cada frase e, finalmente, retira relações gramaticais tais como sujeito-predicado ou objeto-predicado. O último passo caracteriza-se por agrupar termos que contenham relações semelhantes, nas mesmas classes e, em última instância, a ontologia é formada.

Khan e Luo (2002) tiveram como objetivo formalizar uma ontologia de domínio a partir de textos, por recurso a técnicas de clustering e à ontologia WordNet (Beckwith *et al*., 2021). Os documentos a serem analisados são providenciados pelo utilizador do sistema e, uma vez fornecidos, são agrupados por similaridade de conteúdo. Os clusters são hierarquizados através do algoritmo SOTA, sendo atribuído um conceito a cada cluster através de um mecanismo bottom-up. De seguida, cada conceito atribuído é associado ao conceito apropriado e definido na ontologia WordNet.

Conforme referido, existem metodologias que recorrem a técnicas do âmbito da linguística para uma caracterização mais precisa do domínio. Bachimont *et al*. (2002) propuseram um método que se resume essencialmente a três passos. Na primeira etapa, o utilizador escolhe os termos relevantes do domínio, normaliza o seu significado e define as suas características, assim como a hierarquia entre termos, acompanhados por uma justificação que explique a sua posição hierárquica. No segundo passo, há uma fase de formalização de conhecimento que tira partido da taxonomia originada no primeiro passo. Neste momento, é retirada a ambiguidade possivelmente presente nos termos, com a finalidade de um especialista no domínio poder formalizar o conhecimento. Por fim, a taxonomia gerada é transcrita para uma linguagem de representação de conhecimento específica.

Nobécourt (2000) apresentou uma abordagem que tem como principal objetivo construir ontologias a partir de textos com recurso a técnicas de text mining e a um corpus. O método proposto divide-se essencialmente em duas atividades: modelação e representação. A primeira representa uma extração dos termos relevantes do domínio no corpus. Estes termos são modelados como conceitos e constituem o primeiro esqueleto da ontologia. Depois de uma análise efetuada por especialistas do domínio, são identificados os principais subdomínios da ontologia, através dos termos reunidos no passo anterior. Em segunda instância, os conceitos modelados são traduzidos para linguagem natural, o que dá origem a um novo corpus; consequentemente, uma nova lista de primitivas (termos) é gerada, num procedimento iterativo que gradualmente refina a ontologia.

A representação resume-se a uma tradução do esquema de modelação para uma linguagem de implementação, cujo método é tecnologicamente possível através da plataforma TERMINAE (Biébow *et al*., 1999).

Com recurso a técnicas de análise de linguagem natural, Kietz *et al*. (2000) propuseram um método genérico para criar uma ontologia de domínio através de diferentes fontes heterogéneas. À semelhança de diversas metodologias conhecidas, esta também assenta na idea de extrair ontologias a partir de uma ontologia base, sendo a última enriquecida à medida que se identifiquem novos conceitos no domínio. Esta abordagem é semiautomática, de certa forma, porque o utilizador representa uma parte ativa no processo. O papel por ele desempenhado é o de especificar que documentos deverão efetivamente ser usados para enriquecer a ontologia. Todos os novos conceitos são identificados através de técnicas de processamento de linguagem natural e, uma vez enriquecida, a ontologia base é transcrita a um nível específico do domínio, com os conceitos genéricos removidos através de abordagens estatísticas. As relações entre conceitos são extraídas, e caracterizadas, através de métodos de aprendizagem e adicionadas à ontologia resultante. Uma vez que esta é ciclicamente modificada, todo este processo se torna iterativo também. Aussenac-Gilles *et al*. (2000) indicaram uma forma de criar um modelo de domínio através de ferramentas de NLP e de técnicas de linguística para analisar documentos. Este método inicia-se a partir de um texto e, além disso, pode também tirar proveito de ontologias existentes ou utilizar recursos terminológicos para construir uma nova ontologia. Todo o processo está arquitetado em três níveis: a nível linguístico, a nível de normalização e a nível formal. O primeiro é composto pelos termos e padrões léxicos extraídos do texto. Neste passo, é evidenciada a importância do auxílio de um especialista do domínio para a seleção do conjunto de documentos candidatos para análise e, ainda, o estudo e seleção de ferramentas linguísticas adequadas para a análise dos textos.

Uma vez extraídos, os termos são agrupados e traduzidos em conceitos inerentes ao domínio e respetivas relações semânticas, no segundo nível. O nível de normalização apresenta-se dividido em duas subfases: linguística e conceptual. Na primeira subfase são escolhidos os termos e relações léxicas a serem modelados e, para cada termo, é adicionada uma definição em linguagem natural por cada conotação, salvaguardando apenas os significados mais relevantes para o domínio. Durante a subfase conceptual são definidos os conceitos e as relações semânticas numa forma normalizada. Finalmente, identificados os conceitos e relações, estes são descritos, no último nível, através de uma linguagem formal. A avaliação e implementação da ontologia são validadas neste momento e, além disso, todo o conhecimento adquirido é avaliado pelo utilizador e pelo especialista do domínio. Após esta avaliação, a ontologia encontra-se preparada para ser implementada.

Marti Hearst (1992) desenvolveu um método cujo objetivo principal era obter automaticamente as relações léxicas de hipónimos/hiperónimos de um conjunto de documentos, cuja finalidade era a geração de um dicionário de sinónimos de domínio geral. Esta ideia explora um conjunto de padrões léxico-sintáticos predefinidos que, em geral, são facilmente identificáveis. Todo este processo divide-se essencialmente em cinco etapas:

- Escolher uma relação léxica de interesse.
- Reunir uma lista de termos conhecidos por albergarem este tipo de relação.
- Descobrir documentos contidos na obra em que estas relações apareçam como um padrão sintático definido no primeiro passo e registar o ambiente em que os mesmos ocorrem. O ambiente diz respeito ao espaço linguístico no qual estas expressões aparecem, e podem ser excertos de uma frase ou mesmo uma frase completa. Estes registos são reveladores de relações de hiperonímia/hiponímia entre os elementos que a expressão une.
- Descobrir semelhanças entre os ambientes extraídos e a hipótese inicial
- Utilizar um novo padrão, que entretanto tenha sido identificado, para extrair novas instâncias da relação escolhida no primeiro passo, recomeçando o processo no segundo.

Com recurso à ontologia WordNet (Beckwith *et al.*, 2021), é possível validar as relações de hiperonímia/hiponímia se, por exemplo, dois termos apresentarem a mesma ligação hierárquica tanto na ontologia como no dicionário.

Baseando-se na área da semântica distributiva, Alfonseca e Manadhar (2002) sugeriram um método que explora a frequência com que aparecem as palavras associadas a um conceito. A área associada visa quantificar e caracterizar semelhanças semânticas entre palavras ou expressões linguísticas com recurso à sua distribuição em grandes amostras de documentos. O contexto de um conceito pode ser representado por um vetor de palavras, do mesmo contexto, que se evidenciem sob a forma desse mesmo conceito, e as respetivas frequências com que surgem nos registos a analisar. Após a conceção destes vetores, um algoritmo de análise distributiva é aplicado com o objetivo de atribuir níveis de relevância a cada palavra, para que esta seja ou não aceite como candidata a entidade representativa de um conceito. Estes vetores são, ainda, comparados com vetores associados a ontologias existentes, como a WordNet, com o intuito de validar a identificação de hiperónimos candidatos.

2.5 SISTEMAS E APLICAÇÕES

A formalização de uma ontologia é muitas vezes um processo demasiado complexo para se fazer através de definições lógicas, escritas formalmente. Para colmatar a necessidade de simplificar todo o processo de criação e evolução, foram desenvolvidas ferramentas capazes de expressar esta formalização de forma organizada. Vejamos algumas delas.

Protégé

Para a construção de ontologias, o Protégé — Stanford Center for Biomedical Informatics Research, (s.d) — permite estruturar uma ontologia sobre os vários componentes da mesma. Esta ferramenta oferece uma interface gráfica que permite construir manualmente classes, atributos e relações, assim como regras e restrições aplicadas aos componentes referidos. O Protégé (Figura 2) permite criar hierarquia de classes com o intuito de explicitamente evidenciar a taxonomia nos conceitos. Uma vez definidas as classes, as relações e propriedades podem ser definidas e atribuídas de forma simples e intuitiva, e pode-se, inclusivamente, definir o tipo de dados de cada propriedade. É permitido ainda instanciar objetos para povoar a base de conhecimento e exportar toda a ontologia em diversos formatos.

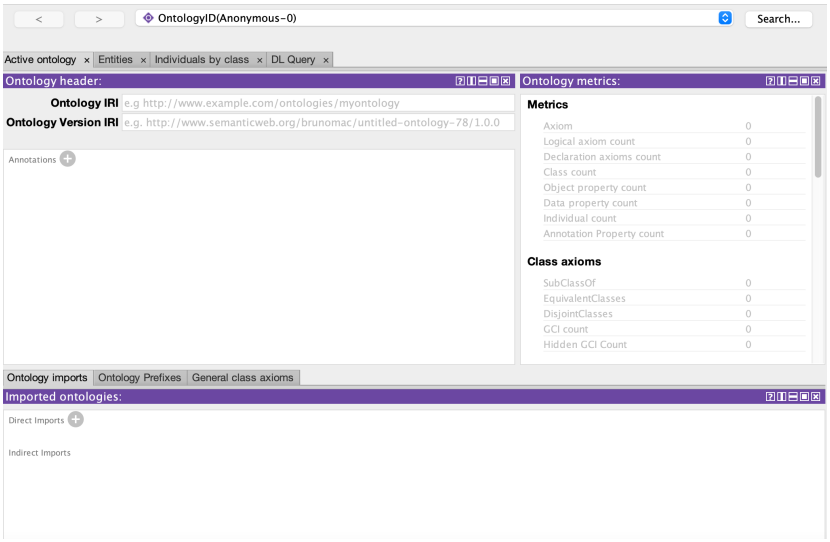

Figura 2: Protégé

SOBA (SmartWeb Ontology-based Annotation)

Desenvolvido na universidade de Kalshruhe (Alemanha), o SOBA (Buitellar *et al*, 2006) povoa automaticamente uma base de conhecimento a partir de informação extraída de páginas web. É composto por um web crawler, componentes para anotação linguística e um módulo final para as transformar em representação ontológica. A temática é o futebol, mais concretamente, o Mundial de 2002 e 2006, e a informação encontra-se representada ou em formato tabular ou em imagens, para ser mais específico, as legendas da imagem.

OntoLearn

Baseado em técnicas de NLP e *machine learning*, o *OntoLearn* (Navigli *et al*., 2004), um sistema de extração automática de ontologias, caracteriza-se por três fases, nomeadamente:

extração de termos, interpretação semântica dos mesmos e evolução da ontologia. Na primeira fase, os termos, que podem ser palavras ou expressões, são extraídos com recurso a técnicas estatísticas de processamento de linguagem natural. Através de documentos genéricos e documentos específicos do domínio, o sistema consegue filtrar termos que não pertençam à terminologia do domínio. Após a inserção dos documentos pertencentes ao domínio no sistema, este último retorna uma lista de termos plausíveis após uma análise sintática. A relevância dos termos é mensurada através da relação direta que um termo tem com o documento e através da aferição dos termos que pertencem ao domínio. Com recurso a estas medidas, os termos candidatos são filtrados através de um grau de associação que determina se o termo é ou não relevante.

A segunda etapa resume-se à análise de palavras compostas ou expressões e à sua correta interpretação. Através de um algoritmo SSI (*Structural Semantic Interconnection*), baseado em correspondência de padrões sintáticos, o sistema executa a tarefa de desambiguação semântica. Este algoritmo produz um grafo semântico que valida a relação entre termos.

Uma vez que os termos tenham sido semanticamente interpretados, a última fase encarrega-se de inserir os novos termos nos nodos apropriados da ontologia inicial. Em última instância, a ontologia gerada é convertida para o formato OWL.

Ontogen

O sistema *Ontogen* (Fortuna *et al.*, 2007) é caracterizado como um editor semiautomático de ontologias que integra algoritmos de machine learning e text mining; o *Ontogen* extrai palavras-chave e sugere-as como conceitos candidatos, a partir de documentos inseridos no sistema como input. Cada conceito sugerido é aceite, ou não, pelo utilizador e, no primeiro caso, anexado ao documento em que foi encontrado. Uma vez efetuado este mapeamento, o utilizador pode procurar por documentos que tenham um determinado conceito. O primeiro protótipo foi utilizado em domínios como negócios, legislações e bibliotecas digitais, e manuseado por especialistas do domínio com pouco conhecimento ou experiência em construção ontológica.

OntoLT

O *OntoLT* (Mylopoulos, 2003) é um plugin do *Protégé* que permite extrair automaticamente conceitos e relações com origem numa coleção de textos linguisticamente anotados. Para cumprir este propósito, o *OntoLT* mapeia as entidades linguísticas contidas nos textos com as classes/propriedades no *Protégé*, com recurso a regras. Os textos analisados, em formato XML, incluem anotações sintáticas e morfológicas, que, consequentemente, permitem relacionar e construir classes, subclasses e relações de um domínio específico. O mapeamento é feito através de expressões X-Path, que se caracterizam como expressões que permitem navegar pela estrutura do XML e extrair os elementos pretendidos. A construção de

regras sintáticas é feita através de algoritmos estatísticos, com o intuito de definir a relevância de cada termo.

Ontobuilder

O *Ontobuilder* (Roitman & Gal, 2006) é uma ferramenta que foi especialmente desenhada para se comportar como um web browser, auxiliando na construção ontológica com informação proveniente de dados semiestruturados, como XML ou HTML. Ao introduzir o URL de uma página web desejada, o seu conteúdo é extraído e a ontologia é formada a partir desse conteúdo. O processo de construção ontológica tem duas fases: uma fase de treino, na qual a ontologia é criada com base no conteúdo extraído, proveniente da página web endereçada no URL apresentado pelo utilizador, e uma fase de adaptação, que se define pelo gradual refinamento da ontologia previamente formada, através de novas extrações de páginas web. Por cada página web extraída, a ontologia resultante desse processo converge com a ontologia já existente.

Text to Onto

O sistema *Text to Onto* (Maedche & Staab, 2000) integra uma infraestrutura de gestão de ontologias com um conjunto de ferramentas capaz de construir uma ontologia a partir de uma existente e mais genérica. Com recurso a regras do domínio e técnicas de *machine learning* para descobrir conceitos, os termos são extraídos conforme a sua frequência de ocorrência e distribuição pelos textos. As relações semânticas e taxonomia são extraídas com recurso a regras de associação ou padrões linguísticos. O resultado produzido traduz-se numa ontologia de domínio (Maedche e Staab, 2000).

3

O CASO DE ESTUDO

3.1 APRESENTAÇÃO GERAL

O objeto de estudo desta dissertação é um conjunto de receitas presente na obra —*"As Receitas de um Frade Português no séc. XVI"* (Barros, 2013). Estas marcas do passado, mas que permanecem bem presentes, foram registadas no manuscrito 142 do Arquivo Distrital de Braga, de origem desconhecida mas que passou pela antiga Biblioteca do Mosteiro de Tibães, encerrando algumas das receitas referências históricas que remontam ao século XVI. A ausência de uma referência ao autor do manuscrito, por falta da primeira folha do mesmo, e o facto de (apenas) ser mencionado num catálogo de Tibães, que data do ano de 1743, como Arte de Dozinha ou Methodo de fazer guizados, s.d., bem como alguns aspetos do seu conteúdo, não confirmam a ideia de que se esteja perante uma arte da cozinha conventual beneditina. O frade José Joaquim de Santa Teresa decidiu, naquela data, reunir variados manuscritos espalhados pela congregação e procedeu ao registo dos mesmos na biblioteca do Mosteiro de S. Martinho de Tibães. Com base nas primeiras gramáticas portuguesas, é possível concluir que a língua e as grafias variáveis presentes na obra aparentam remontar aos séculos XVI-XVII, e daí o título atribuído à obra em estudo, na qual se edita pela primeira vez o manuscrito, anónimo e sem data, mas com referências históricas quinhentistas no interior de várias receitas (Barros, 2013).

O primeiro dos três cadernos do manuscrito 142 apresenta uma vasta coletânea de pratos de culinária, agregando aromas de várias culturas e civilizações, que culmina numa mistura exótica e muito característica. A obra *"As Receitas de um Frade Português no séc. XVI"* dá a conhecer uma fonte rica em matéria de estudo, com diversos elementos, que, apesar de se repetirem em diversas receitas, não traduzem um resultado análogo. São ainda apresentadas algumas anotações na edição do manuscrito, de índole diferente na lição semidiplomática (que inclui toda a variação ortográfica) e na interpretativa (com grafia atualizada), para tornar a sua leitura e compreensão ainda mais precisas.

A obra contém quase trezentas receitas, e um diversificado conjunto de combinações de processos, que dão origem a diferentes resultados. A sequência e repetição destes processos são fatores a ter em conta na análise dos diversos documentos, bem como os

ingredientes e respetivas quantidades. No âmbito desta dissertação, serão exploradas as diversas indicações culinárias, cuidadosamente identificadas e com grafia modernizadas, o que permitirá extrair fatores relevantes sobre as mesmas. Após serem meticulosamente analisadas, essas indicações serão reunidas a outras existentes e respetivas propriedades. Depois, cada receita será incorporada num ficheiro de texto, para que, em fase posterior, possam ser analisadas individualmente. Cada registo de cariz culinário encontra-se escrito de forma peculiar, num português erudito e pouco convencional. Porém, é fácil a sua compreensão, uma vez que é bastante similar ao dos dias de hoje, e, escrito desta forma, preserva a autenticidade de cada receita.

3.2 PROBLEMAS E DESAFIOS

Esta dissertação assenta no estudo de um manuscrito de receitas, editado em lição conservadora (respeitando a grafia do mesmo), tendo como intuito a extração, a análise e a exposição do seu conteúdo culinário. Os registos de culinária, presentes neste manuscrito, revelam os procedimentos da cozinha portuguesa, assim como especiarias e tradições, datados até ao século XVI. Por forma a avaliar a complexidade e o desenvolvimento de um sistema capaz de cumprir os requisitos dos mecanismos de extração automática de ontologias, foram enfrentados os seguintes problemas e consequentes desafios:

1. **Caracterização da linguagem natural** — Numa primeira fase, é crítico e imperativo o estudo deste manuscrito, não só a nível contextual como a nível estrutural, com o objetivo de estabelecer uma base de conhecimento e perceber os padrões de escrita, assim como caracterizar a linguagem natural. Assim, para dar início ao processo de extração, é necessário fazer uma tradução da linguagem natural com o propósito de caracterizar cada termo.

2. **Processo de extração** — Esta etapa não considera apenas a realização de uma simples escolha de abordagem e, consequentemente, da sua execução. Cada caso de estudo tem as suas particularidades, tais como padrões, linguística, frequência de termos e heterogeneidade entre documentos (já que as receitas são habitualmente colecionadas de várias fontes, ao longo dos anos e séculos, até surgirem juntas num só códice). Para além do estudo destas particularidades, é ainda necessário efetuar alguns testes práticos para provar a sua aplicação. Ainda nesta etapa, e dependendo da abordagem escolhida, devem ser caracterizados padrões de escrita, no caso de uma abordagem bottom-up, definidos mecanismos de controlo de frequência de termos, numa abordagem estatística, ou, numa metodologia top-down, analisado sintaticamente o conteúdo dos documentos.

3. **Caracterização de conceitos** — Ainda na análise do caso de estudo, é importante identificar quais os conceitos que devem ser formalmente identificados. Alguns podem ser identificados simplesmente pela lógica adotada por outros documentos do mesmo domínio, enquanto outros terão que ser identificados através de regras formalmente definidas ou por relações que revelem a sua identidade.

4. **Formalização de regras e relações** — Esta é, provavelmente, a etapa mais decisiva na construção ontológica. As regras e relações não só caracterizam cada termo como homologam cada entidade representativa a um conceito. O desafio inerente à caracterização de regras e relações apresenta-se como um conjunto de caracteristicas e atributos que devem ser descritos formalmente, evidenciando univocamente o lugar de um termo na ontologia.

5. **Aplicabilidade num sistema real** — Aqui pretende-se descobrir como tirar partido da ontologia gerada e tirar proveito da semântica que a mesma oferece. Em que circunstâncias a mesma pode ser modificada e como tornar o seu crescimento o mais escalável e evolutivo possível, são dois aspetos a ter em conta. Em suma, definir como tornar visíveis as vantagens que a ontologia oferece.

Além dos problemas enumerados, neste trabalho foram identificados alguns desafios, no processo de extração, que requerem alguma atenção, sendo eles:

1. **Semântica** — Um dos problemas detetados em primeiro lugar foi o facto de os manuscritos possuírem uma escrita algo erudita, o que em algumas situações torna complicada a correta interpretação do processo ou ação para o analisador sintático. Contudo, este problema deverá ser ultrapassado se as regras definidas para cada conceito o caracterizarem corretamente.

2. **Análise sintática** — Tendo em conta o registo de escrita presente nos textos analisados, por vezes é quase impossível identificar processos que se encontrem sob a forma de substantivos. Tendencialmente, os verbos são sempre traduzidos em ações e é com base nessa lógica que a extração de processos se baseia, em grande parte dos casos.

3. **Processos de culinária** — Provavelmente, a tarefa mais complexa neste caso de estudo centra-se na distinção entre uma simples ação e um processo estritamente ligado à culinária. Por exemplo, a frase "... lhe metem dentro um ou dois olhos de couves segundo a quantidade de sopas que querem **fazer**, ...", não nos permite distinguir formalmente se o verbo anotado é um processo de culinária ou uma ação que não seja propriamente reveladora do domínio em estudo. Neste caso, o verbo anotado poderia facilmente ser substituído pelo verbo "ferver", que já se enquadra num processo ligado à culinária.

4. **Precisão** — Em condições ideais, o sistema deveria conseguir identificar e caracterizar todos os processos de culinária presentes na confeção de uma receita. Contudo, tal não acontece, devido ao facto de cada documento ter a sua própria identidade e cada processo poder apresentar-se de forma quase impossível de ser identificada. No entanto, um sistema que apresente uma taxa de precisão acima dos 80% disponibiliza-nos resultados bastante aceitáveis e viáveis.

3.3 IDEALIZAÇÃO DA ONTOLOGIA

Após a análise dos textos de culinária, foi possível dividir cada receita em três blocos de informação (Figura 3) para facilitar o tratamento dos dados, podendo cada um ser caracterizado da seguinte forma:

- Bloco 1 — Constituído por identificador da receita e respetivo título.
- Bloco 2 — Caracterizado pela descrição dos ingredientes pertencentes à receita, e opcionalmente, podendo conter algumas anotações adicionais — o conteúdo deste bloco II (Ingredientes) não figura no manuscrito, sendo da responsabilidade do editor (Barros, 2013), que extraiu esses dados do corpo de cada receita, onde figuravam em qualquer posição, sempre variável, e os apresentou em ortografia contemporânea, para facilitar a consulta e utilização das mesmas.
- Bloco 3 — Representa a descrição do processo de preparação de uma receita.

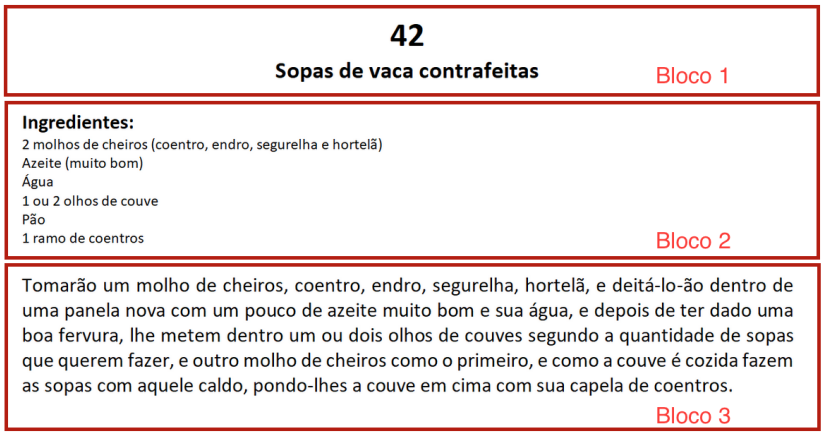

Figura 3: Estrutura básica de uma receita

A definição das classes pertinentes na ontologia de culinária depende altamente das unidades representativas que uma receita apresenta. Neste trabalho, foram reconhecidas seis classes que permitem incorporar as entidades presentes em cada documento, sendo elas:

- **"Receita"** — Classe que representa cada um dos registos culinários ou textos para confeção de um prato e as suas características gerais. Uma receita é caracterizada pelo seguinte conjunto de atributos: “identificador”, “título” e “data”. O valor do atributo “data” não é extraído do documento, mas sim atribuído pelo sistema no momento de inserção da ontologia.
- **"Ingrediente"** — Designação de um ingrediente e respetivas notas adicionais. Os atributos desta classe – "nome", "quantidade", "alternativas"e "observações"–, à exceção do nome, nem sempre serão preenchidos, devido à inexistência de unidades que evidenciem quantidades, alternativas a um ingrediente ou até mesmo observações.
- **"Nota"** — Classe que nem sempre irá estar presente num documento, pois representa uma anotação culinária, da mão do editor (que a retirou ou deduziu do texto da receita), inserida no bloco dos ingredientes. Cada nota terá uma designação – todo o texto que representa a anotação até ao sinal de pontuação "– e o conteúdo são todas as linhas que estão indentadas imediatamente após a designação, ou, em certos casos, todas as linhas até à próxima nota.
- **"Procedimento"** — É a unidade que representa todas as etapas da confeção do prato identificado pela receita. Define-se como *designação* a propriedade que contém toda a preparação da receita.
- **"Processo"** — É uma classe que representa cada um dos processos identificados no bloco da preparação. As entidades que representam esta classe na ontologia aparecem de modos variados quanto à sua forma verbal, género ou número. Estas formas são consideradas como variantes, e a sua forma primitiva como nome.
- **"Indice"** — Representa a enumeração de todas as palavras presentes no documento. Cada registo inclui o nome da palavra, permitindo a soma dos mesmos permite saber o tamanho textual das receitas introduzidas.

3.4 O PROCESSO DE EXTRAÇÃO

3.4.1 *Ontology Learning Layer Cake*

No que diz respeito à evolução de ontologias a partir de textos, a abordagem mais conhecida é a *Ontology Learning Layer Cake* (Mishra & Jain, 2014). Esta abordagem permite dividir o processo de extração em quatro ou cinco etapas (Figura 4), sendo cada uma delas definida de acordo com a área de domínio. Na primeira etapa, são extraídos todos os termos presentes no texto, com o objetivo de criar um conjunto de termos prontos a serem catalogados. Na fase seguinte, cada termo é agrupado com todos os sinónimos encontrados no conjunto anterior,

o que facilita a identificação dos conceitos que os mesmos representam. A terceira etapa associa os termos a um conceito e infere, pela primeira vez, uma correspondência entre a linguagem do domínio e o conceito ontológico. O quarto degrau da pirâmide representativa desta abordagem assenta na inferência hierárquica de conceitos que representam um conceito mais básico, como, por exemplo, um professor ser uma pessoa. No quinto passo, são definidas as relações entre conceitos e, por sua vez, especifica o domínio e alcance da relação. Na derradeira etapa, é construído um conjunto de regras que permite inferir o conceito e respetivas ligações semânticas, baseado em características inerentes ao conceito e suas propriedades.

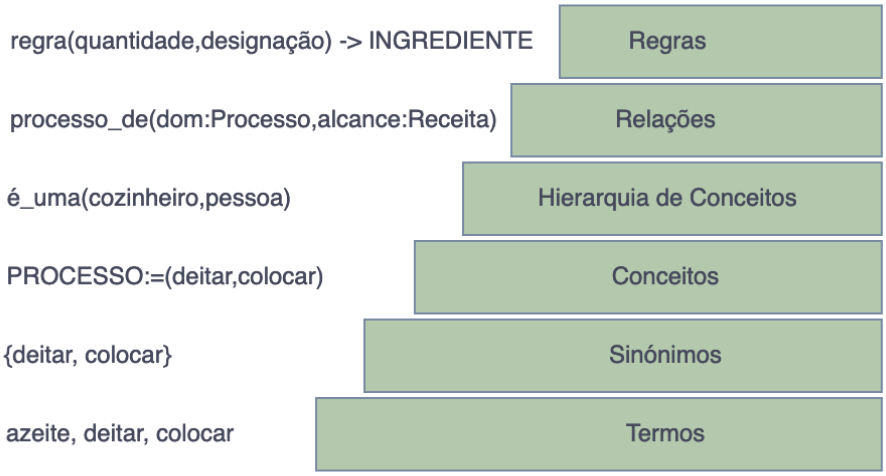

Figura 4: *Ontology Learning Layer Cake* (Mishra & Jain, 2014)

3.4.2 *Processamento Textual*

Para iniciar o processo de extração é necessário um pré-processamento do texto de uma receita, nomeadamente dos três blocos referidos anteriormente (Secção 3.3). Para tal, o uso de expressões regulares é vital relativamente ao processamento de linguagens, e, neste sentido, foram definidos os seguintes padrões textuais:

- `[0-9]+\n` que permite extrair o identificador da receita, pois o mesmo é o primeiro elemento a aparecer no documento. Como esta propriedade é fundamental para controlo de documentos repetidos, cada documento deverá, imperativamente, conter este elemento.
- `[0-9]+\n.+[\^\n]` que permite extrair tanto o identificador como o título da receita. O prefixo desta expressão regular caracteriza-se pela expressão mencionada para extração do identificador. O prefixo é a parte responsável pela extração do título da receita. O título da receita encontra-se sempre na linha abaixo do identificador.
- `Ingredientes\:\n.*(?:\r?\n(?!\r?\n).*)*` que é a expressão regular responsável pela extração dos elementos do bloco 2. Para facilitar a leitura e interpretação da mesma, pode-se olhar para este padrão da seguinte forma:
  - `Ingredientes\:\n` que é responsável pela captura do cabeçalho dos Ingredientes.
  - `.*` que unifica com todos os carateres presentes numa linha textual.
  - `(?:` que define o início de um padrão que envolve um conjunto de carateres que não se pretende capturar.
  - `\r?\n` que representa qualquer indentação ou quebra de linha após uma linha textual.
  - `(?!\r?\n)` que representa um lookahead negativo, que pretende dizer que não pode existir mais nenhum tipo de indentação ou quebra de linha para além do que foi extraído pela expressão anterior. Este padrão é justificado com a divisão entre o bloco 2 e o bloco 3, que é feita, pelo menos, por duas quebras de linha.
  - `.*)*` que extrai todos os carateres, excepto quebras de linha, zero ou mais vezes. Juntamente com os três padrões anteriormente citados, toda a expressão contida nos parênteses mais externos é capturada zero ou mais vezes igualmente.
- `\n{2,}` que procura no texto todos os conjuntos separados por, pelo menos, duas quebras de linha e extrai o último. Utiliza-se esta abordagem porque o bloco 3 não tem qualquer cabeçalho ou outro padrão restritivo.

Todos estes padrões foram retirados com recurso a funções do módulo *re* do Python (re — Regular expression operations — Python 3.10.5 documentation, s.d). Este módulo disponibiliza um conjunto de operações que permitem o uso de expressões regulares para processamento de dados. No âmbito deste trabalho foram utilizadas funções como o *findall* e *finditer*, para extrair conjuntos textuais que obedeçam ao padrão estabelecido.

3.4.3 *Extração de Termos*

Com base na ontologia idealizada, os termos que requerem mais processamento são os ingredientes e os processos, pois exigem uma compreensão do significado na forma variável em que estão redigidos no título e no corpo da receita (pois na secção Ingredientes estes acham-se reunidos e apresentados em grafia atualizada pelo editor). Para extrair os termos compreendidos no identificador, no título da receita e na preparação, foi simplesmente usado um processo de divisão por tokens. Esta segmentação foi efetuada com o auxílio da biblioteca *Spacy*, com o módulo `pt_core_news_lg`, que mapeia todas as palavras extraídas do texto com base num conjunto de dados previamente treinados. Além desta biblioteca, pode-se, opcionalmente, efetuar a divisão por tokens através de expressões regulares.

Cada ingrediente presente no documento é tratado linha a linha. Contudo, cada uma delas pode referir um ou mais ingredientes. Numa primeira fase, a linha é desdobrada em elementos singulares, caso haja uma enumeração de elementos. Para obter estas partições, foi utilizada uma expressão regular que permite dividir cada linha através de vírgulas ou pela conjunção *e*. Tal padrão caracteriza-se pela expressão `\se\s+(?![^()]*\))|,\s+(?![^()]*\))` e pode ser explicado da seguinte forma:

- `\se\s+` — que caracteriza a conjunção *e*, usada para enumerar dois ingredientes, que é antecedida por um espaço em branco e precede um ou mais espaços em branco.
- `(?![^()]*\)` — que verifica se aquilo que sucede à conjunção não é uma anotação compreendida entre parênteses (*lookahead negativo*), pois se tal acontece, é desconsiderado como ingrediente e aceite como uma observação.
- `|` — que remete para a existência de uma alternativa ao padrão referido anteriormente.
- `,\s+(?![^()]*\)` — que extrai dois ingredientes separados por vírgulas, aplicando a mesma lógica das duas expressões iniciais, com a diferença de que a conjunção *e* é substituída por uma vírgula.

Uma vez aplicado este filtro no bloco dos ingredientes, obtém-se um conjunto de termos que posteriormente serão associados a conceitos, conforme as relações e regras inferenciadas.

Relativamente ao bloco de texto referente à preparação da receita — a original do manuscrito —, é apenas efetuada um mapeamento de todas as palavras presentes neste texto. Além deste processo, começa aqui um processo que é invisível ao utilizador e que confere um grau de automatismo elevado ao sistema: a extração de potenciais termos relevantes ao processo de preparação da receita.

3.4.4 *Padrões de Hearst*

Cada documento textual possui as suas características, que variam consoante fatores temporais, externos e pessoais. Assim como o ser humano tem impressões digitais, a pessoa que redige os textos tem a sua própria identidade incorporada na escrita. Para reconhecer esta assinatura, é necessário um pequeno estudo com a finalidade de aprender a forma como as frases são articuladas, e, sintaticamente, o modo como as mesmas são construídas.

Os padrões de Hearst são padrões sintáticos e lexicais de alta precisão, que surgiram da necessidade de detetar hiperónimos de um determinado contexto. Estes padrões são definidos com o propósito de identificar termos que possam ser relevantes. Quanto mais específicos e direcionados forem esses padrões para o relevo semântico contido no domínio, melhores resultados serão produzidos. Relativamente à temática de *Ontology Learning*, estes padrões auxiliam na produção de regras e relações entre conceitos, que por sua vez facilitam na recolha de dados.

Na prática, para extrair os elementos sintáticos e lexicais presentes na escrita do autor, é definido um conjunto de padrões, apoiado no estudo das várias receitas contidas na obra. Cada padrão identifica marcas linguísticas utilizadas pelo autor, na forma de termos, que, posteriormente, serão avaliadas como aceitáveis ou sem interesse aquando da sua inserção no sistema ontológico. Neste trabalho foram identificados dez padrões, uns mais genéricos e outros mais específicos. Nem sempre os moldes mais específicos cobrem todos os casos. Definiram-se estes padrões da seguinte forma:

1. `[^\s]+-[^\s\.]*` — Esta expressão regular compõe o padrão que identifica todos os verbos com pronomes associados. Durante a análise sintática do procedimento da receita, observou-se que, geralmente, um verbo seguido de um pronome enclítico (pós-verbal) é um processo culinário. Neste caso, estes verbos surgem essencialmente no início das frases. No entanto, esta particularidade não está contemplada na expressão, para que a extração seja mais permissiva.

2. `[^\s,\.]+ [^\s,\.]+` — Este padrão extrai duas palavras: uma contida nos ingredientes e outra relativa ao estado do ingrediente. Este estado será interpretado como um verbo na forma de particípio.

3. `[^\s,\.]+ com` — Que identifica um verbo seguido da preposição *com*. Na culinária, geralmente a preposição *com* está associada a processos que estão diretamente relacionados com um ingrediente.

4. `que [^\s,\.]+ [^\s,\.]+|se [^\s,\.]+ [^\s,\.]+|lhe [^\s,\.]+ [^\s,\.]+` — Que representa um excerto da frase que comece com um pronome, seguido de uma outra palavra de qualquer foro sintático, e finalizadas com um verbo que possa indicar um processo. Os pronomes *que*, *se* e *lhe* indicam, na escrita destas receitas, uma sucessão de um possível processo culinário.

5. `se [^\s,\.]+ com [^\s,\.]+` — Este é um padrão particular, que se repete algumas vezes em vários documentos, começando com o pronome 'se', seguido de um possível processo, a que se segue a preposição 'com' e uma palavra que se assume como um ingrediente. Este padrão é dos mais eficazes por ser tão particular, dele resultando termos bastante plausíveis a serem aceites como processos.

6. `[^\s,\.]+ em [^\s,\.]+` — Que identifica um processo que antecede a preposição 'em' e uma palavra que represente um ingrediente. O facto de existir um ingrediente precedido da preposição 'em' indica que alguma ação será efetuada sobre o mesmo, logo, será plausível considerar a primeira palavra como candidato a processo.

7. `se [^\s,\.]+| lhe [^\s,\.]+` — Que poderá identificar pronomes seguidos de verbos. Estes verbos poderão ser processos, no entanto, por ser uma característica tão genérica, é provável que sejam extraídos verbos pouco alusivos a processos.

8. `[^\s,\.]+ [^\s,\.]+` — Que identifica duas palavras, à semelhança do segundo padrão, no entanto, é também usado com um outro intuito. Estas duas palavras caracterizam um verbo composto, sendo a primeira o verbo auxiliar e a última o verbo principal. O interesse deste padrão reside na segunda palavra, dado que, nesta obra, é recorrente o uso de verbos compostos ou perífrases verbais, e que, por sinal, poderão indicar um processo.

9. `[^\s,\.]+` — Que captura todas as palavras existentes na preparação e verifica, uma a uma, se já são um processo identificado numa importação anterior. Esta expressão acrescenta um fator evolutivo ao sistema, o que se traduz numa melhor e mais eficaz inserção de processos na ontologia.

10. `e [^\s,]+` — Este é um padrão genérico, que identifica a conjunção 'e' com um verbo que se assume como candidato a processo. Aliado a esta expressão, está um baixo fator de plausibilidade, mas que por vezes se traduz numa extração correta.

	Frase	Processo obtido	Receita
Padrão 1	"Deitar-se-á em um tachinho..."	deitar	145
Padrão 2	"...salsa e cebola picados,..."	picar	96
Padrão 3	"...e clarificado com uma clara de ovo..."	clarificar	145
Padrão 4	"...a qual se lhe lançará muito bem..."	lançar	145
Padrão 5	"...e nela se frigirão com azeite..."	frigir	94
Padrão 6	"...e afogados em azeite se lançará..."	afogar	87
Padrão 7	"...um fogo brando se derreterá nela..."	derreter	145
Padrão 8	"...havendo primeiro sido passadas..."	passar	145
Padrão 10	"...passadas e aboboradas pelo açúcar..."	aboborar	145

Tabela 1: Exemplos de aplicação de Padrões de Hearst

Na Tabela 1 podemos ver alguns dos padrões extraídos de diferentes receitas, com os respetivos processos obtidos. De referir que o padrão 9 depende apenas dos processos existentes no sistema

Após a extração dos termos caracterizados pelos padrões de Hearst anteriormente descritos (Tabela 1), é necessário classificar sintaticamente o conteúdo da preparação de uma receita. Para tal, o texto relativo ao procedimento culinário para a confeção da receita é introduzido num tokenizer, fornecido pela biblioteca *Spacy*, que foi adaptada à língua portuguesa. Assim que todas as frases estejam devidamente classificadas, é possível extrair a função sintática atribuída a cada palavra, vulgo *token*. Esta atribuição revela-se vital para, em conjunto com os padrões estabelecidos, moldar as regras para a aceitação dos termos. Cada palavra é instanciada num objeto, cujas propriedades podem ser o seu significado literal, a redução da palavra à sua raiz e a sua classe gramatical, com especial atenção esta categoria gramatical, que é o atributo que permite identificar os verbos, expressos nas variadas formas verbais.

Os atributos descritos estão disponíveis da seguinte forma:

- token.text — que é a propriedade que devolve o termo conforme extraído.
- token.pos_ — que revela a classe gramatical que o termo encontrado representa.
- token.lemma_ — que representa a palavra primitiva ou raiz da qual o termo extraído surgiu.

Concluído o processo de segmentação por tokens, estão reunidas as condições para aplicar as regras definidas pelos padrões de Hearst. Todos os termos extraídos pelas expressões regulares juntamente com a classe gramatical correta são considerados como aceitáveis.

A cada processo são associadas todas as formas em que este aparece escrito no texto. Na prática, é feito um mapeamento de todas as variantes deste processo relativamente à

sua forma verbal, género e número. Desta forma, é mais fácil identificar as ocorrências de um processo em determinada receita (Figura 5).

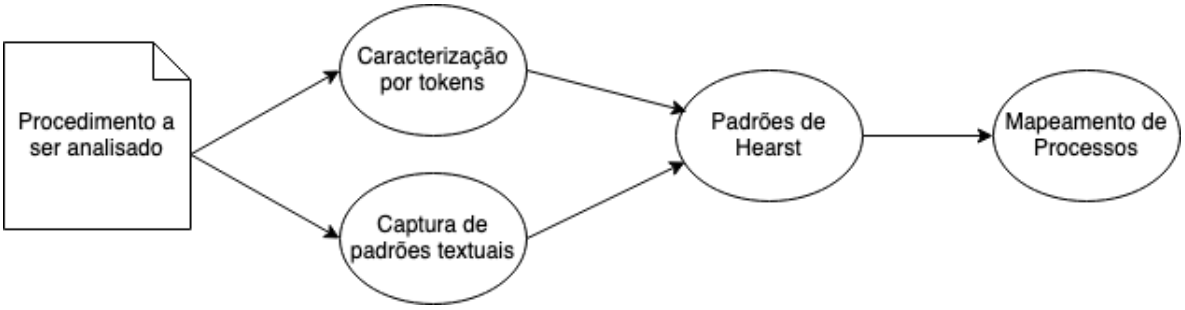

Figura 5: Extração de processos de culinária através dos Padrões de Hearst

3.4.5 *Conceitos*

Esta fase, que se caracteriza como posterior à extração de termos e agregação de sinónimos, prevê a atribuição de uma representação denotativa da classe a que os termos pertencem. Neste trabalho, a atribuição de classes aos termos extraídos é, em certa medida, evidente, uma vez que a sua localização no documento revela, na maior parte das situações, a sua verdadeira identidade. As entidades contidas no bloco A assumem a classe “Receita”, porque, na realidade, este bloco contém sempre os mesmos elementos. A mesma situação já não se repete nos restantes blocos, ou, pelo menos, não de forma tão direta, uma vez que as classes associadas dependem diretamente dos algoritmos de extração de termos e, também, das relações e regras associadas. É neste sentido que a parte automática do sistema começa a ganhar forma, pois a sua identidade vai sendo construída conforme o processamento executado pelo sistema.

No bloco B, vão ser descobertas entidades pertencentes à classe "Ingrediente", no entanto, poderão existir entidades que pertençam à classe "Nota", pois não constituem parte integrante da receita, mas auxiliam, de forma indireta, na elaboração da mesma. Na derradeira secção do documento, é garantido que, para além dos ingredientes, extraídos manualmente pelo editor e por ele colocados no bloco B, surja um elemento representativo da classe "Procedimento", ainda que possam surgir processos derivados do processamento do conteúdo desta classe. Por último, e conforme referido, são identificadas entidades representativas da classe ontológica "Processo". Juntamente com esta identificação, surgem os vários momentos da confeção da receita, uma vez que a preparação de um prato de culinária envolve sempre, pelo menos, um processo. O não retorno de processos identificados poderá induzir uma descrição procedimental pobre ou incompleta.

Relativamente à distribuição hierárquica de conceitos, pode ser encontrada uma evidência da mesma aquando da agregação de processos extraídos da preparação. Qualquer processo identificado é uma abstração de uma definição mais específica, como uma ação realizada pela pessoa a preparar a receita, ou simplesmente por um procedimento meramente exclusivo do mundo da culinária. Como caso concreto, é correto dizer que o processo "deitar"infere uma ação do utilizador e que o processo "ferver"representa uma ação relativa a um ingrediente. Contudo, esta diferenciação não foi explicitamente integrada na ontologia, uma vez que os padrões definidos não estão desenvolvidos para fazer essa distinção hierárquica.

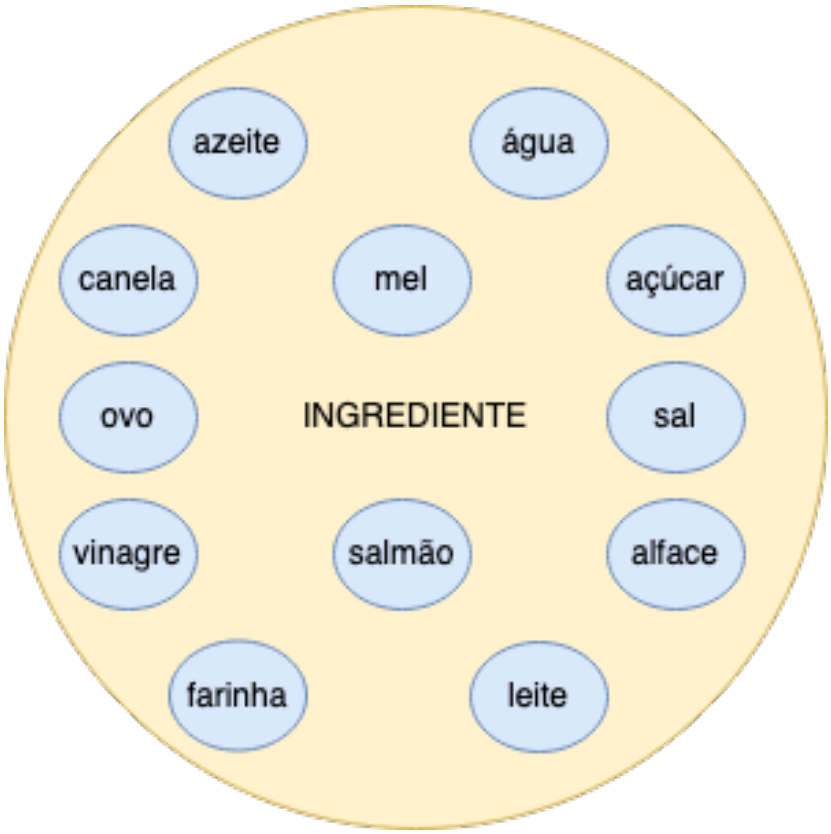

Figura 6: Exemplo de um conceito e respetivas entidades

3.4.6 *Relações e Regras*

A última etapa da metodologia de ontology learning adotada confirma a identidade de cada termo, assim como o papel que o mesmo irá desempenhar na área semântica (Mishra & Jain, (2014). As relações, tal como entre os seres humanos, permitem estabelecer cadeias de conhecimento e afinidade entre duas entidades, em que cada relação une ambas por uma ou mais características. Tal como as propriedades da classe, a relação é parte integrante no processo de determinação do papel que um determinado elemento desempenha na ontologia. Quando não se consegue apurar quais as propriedades de um determinado elemento, seja porque as mesmas não existem ou pelo simples facto de que estas não sejam suficientes para identificarem univocamente a classe a que este elemento pertence, se uma relação for identificada, é plausível o correto reconhecimento da entidade em causa. Cada relação é composta por um domínio e o seu respetivo alcance, normalmente definidos por lógica semântica e a qual desempenha um papel fulcral na interação entre conceitos.

Nas receitas analisadas, todas as ligações entre conceitos foram criadas pelos termos encontrados, e todos estes termos se relacionam com o conceito "Receita". Esta lógica justifica-se pelo facto de todos os elementos encontrados serem parte constituinte de uma receita, e, objetivamente, pela relação direta entre um determinado conceito e todas as receitas introduzidas no sistema. Cada conceito encontrado será relacionado com a receita em análise, na qual a relação assume um grau de posse e de parte integrante. Quando o domínio é "Receita", a relação expressa-se como uma ligação possessiva, e quando esta se caracteriza como o conjunto de chegada, a relação é identificada como uma parte da receita.

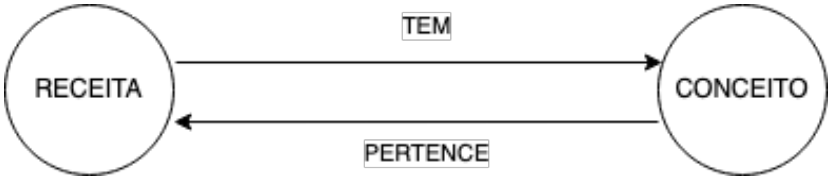

Figura 7: Exemplo da relação entre "Receita"e qualquer conceito da ontologia

As regras determinam se o termo em análise obedece a um conjunto de fatores ou propriedades que caracterizam um conceito na ontologia. Quanto mais detalhe houver nestas regras, menor será a margem de erro na associação dos termos aos conceitos. Cada termo candidato tem que contemplar obrigatoriamente o conjunto de propriedades que definem cada conceito, juntamente com as relações que lhe são inerentes. No entanto, no caso do conceito "Ingrediente", nem todas as propriedades necessitam de ser preenchidas, uma vez que nem sempre estão presentes no documento, com a designação do ingrediente como exceção. Além destas premissas, todo e qualquer campo extraído obedece a um

conjunto de regras morfológicas para que seja considerado um termo, tal como observado nos padrões de Hearst, ou na extração dos ingredientes.

3.5 PRESERVAÇÃO DA ONTOLOGIA

Decorrido o processo de extração, começa a etapa de armazenamento do conhecimento extraído e, nesse aspeto, uma base de dados suportada por grafos cumpre os requisitos de uma ontologia. O Neo4j (Neo4j Graph Data Platform, s.d) é um sistema de gestão de base de dados que respeita o conjunto de propriedades ACID e que, através da linguagem *Cypher*, permite fazer pesquisas nos dados da ontologia. O registo das novas entidades extraídas pode ser feito com uma combinação ou junção (merge) dos dados pré-existentes com o conhecimento a ser importado, ou, caso não existam referências, é inserida toda a informação nova, por completo, na ontologia. Apesar da existência de um nodo do tipo Processo com o mesmo nome, será criado um novo nodo porque este difere nas variantes que este processo tem em cada receita.

O aperfeiçoamento da ontologia é feita depois do último passo da importação do documento e tem dois momentos distintos: 1) a atualização da ontologia contida no Neo4j e 2) a atualização de um ficheiro OWL que também contém a ontologia. A grande diferença entre as duas remete para a questão da inferência de semântica, uma vez que no Neo4j é feita pelas queries definidas e executadas, enquanto no ficheiro OWL é feito por um *reasoner*, o *HermiT* (Glimm et al, 2014). O *reasoner* permite completar qualquer informação que possa não ter sido adjudicada ao conceito, ou, por outras palavras, determinar qual a classe de uma entidade com base numa propriedade, seja esta um atributo ou uma relação. A grande desvantagem deste processo é que é extremamente moroso, principalmente se o cálculo das novas relações, e inferências, for feito após a inserção de múltiplos registos. Esta última serve como justificação para o facto de a manutenção do ficheiro OWL ser feita sempre que um novo documento é importado.

No caso da remoção, quando se pretende eliminar um registo da ontologia, é apagado o nodo da receita pretendida e, por consequência, todas as relações da mesma são eliminadas. Após a perda de ligação semântica com a receita eliminada, as restantes entidades previamente ligadas permanecem no sistema caso integrem relações com outras receitas ou, caso contrário, são também eliminadas, por deixarem de existir no contexto de qualquer receita remanescente. Tal como na inserção de novos registos, o *reasoner* trabalha sobre o ficheiro OWL a cada registo eliminado.

4

ANÁLISE DE RESULTADOS

4.1 O AMBIENTE DE TESTES

Para permitir a divulgação de resultados e a experimentação por parte do utilizador do sistema ontológico produzido, foi concebida e desenvolvida uma plataforma WEB. O desenvolvimento desta plataforma foi feito através da *microframework Flask*, escrita em *Python*. Esta aplicação WEB foi arquitetada como um padrão MVC (*Model-View-Controller*, não só pela organização da aplicação entre a camada lógica e a camada gráfica, mas também pela facilidade em acrescentar novas funcionalidades e modificar existentes. Relativamente à lógica da aplicação, para além de métodos de processamento de informação, são também definidas operações sobre a base de dados em *Neo4J*. Este sistema de gestão de base de dados orientada para grafos foi escolhido pela sua capacidade de representar uma ontologia e todas as receitas importadas no sistema são armazenadas nele. Para além do *Neo4J*, a persistência dos dados é feita também num ficheiro OWL, com o objetivo de este ser descarregado pelo utilizador, na plataforma. Toda esta camada lógica alimenta a interface do utilizador, que foi desenvolvida com recurso à *framework* de *front-end Bootstrap*.

A plataforma multiutilizador permite a integração de receitas, mas não garante a vinculação dos documentos introduzidos ao utilizador que os introduziu. Todos os utilizadores terão acesso aos textos de cariz culinário, desde que se encontrem registados na plataforma e com permissão para adicionar ou eliminar receitas do sistema. Após a autenticação (Figura 8) é apresentado um painel de visualização de dados, ou *dashboard*, que promove a apresentação de diversas características do sistema ontológico.

Figura 8: Autenticação no sistema

A aplicação dispõe de um conjunto de funcionalidades que permitem a cada utilizador uma experiência mais imersiva e prática com o sistema de extração automática de ontologias. Este conjunto de funcionalidades é constituído por:

- **Home** — (Figura 9) — É a primeira página que o utilizador encontra após a sua autenticação no sistema. Aqui poderá ver o número de receitas inseridas, a soma de palavras indexadas, a quantidade de ingredientes e, ainda, a totalidade de processos extraídos. Esta página é meramente informativa, não havendo qualquer interação com o utilizador, para além da exposição de informação relativa aos documentos registados.

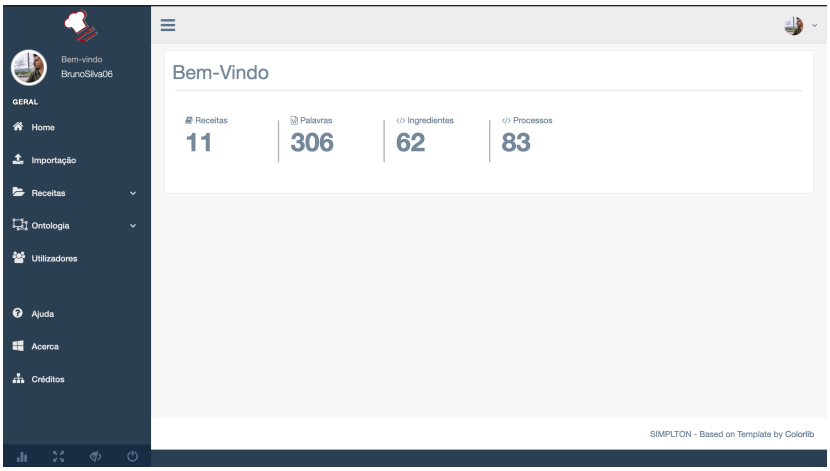

Figura 9: Ambiente de entrada — *Home* — do sistema

- **Importação** (Figura 10) – Esta é a funcionalidade mais importante do sistema, uma vez que é através dela que a lógica e semântica do mesmo “ganham vida”. No processo de importação dos textos de uma receita, são dadas duas escolhas ao utilizador: a

inserção da receita por campos, ou inserção da receita por documento. Na primeira opção o sistema apresenta ao utilizador um formulário para que possa introduzir a informação relativa aos seguintes elementos de dados: “ID Receita”, “Nome da Receita”, “Ingredientes”, "Preparação"e, opcionalmente, "Adicionar Foto do Texto Manuscrito da Receita". Na segunda opção, o utilizador tem o trabalho de inserção mais simplificado, uma vez que apenas tem que inserir o documento, conforme especificado anteriormente (ver secção 3.3), após o que o sistema irá extrair as várias unidades estruturais da receita automaticamente.

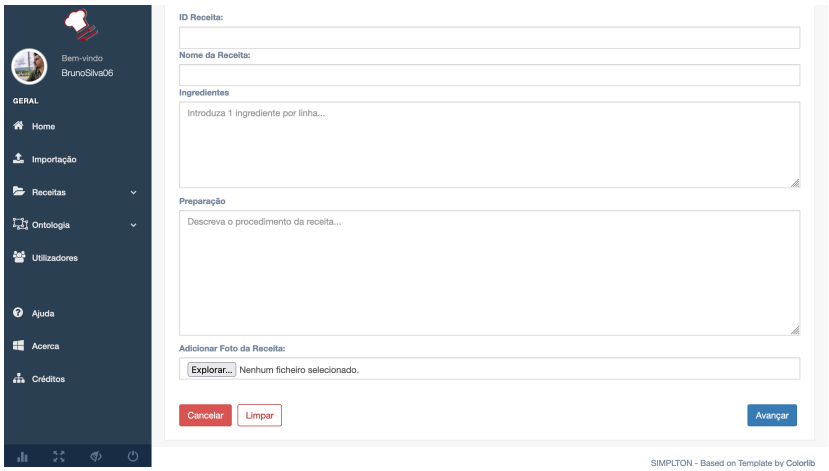

Figura 10: Importação de receitas — o primeiro passo.

No passo seguinte (Figura 11), o utilizador tem oportunidade de observar e analisar informação pertinente nas diversas componentes da receita:

- **Ingredientes** — No qual é apresentada informação, em formato tabular, relativa à designação do ingrediente, bem como a quantidade e a unidade de medida, observações e alternativas. O preenchimento de todas as colunas é feito com os dados disponíveis na receita. Além disso, é possível extrair também anotações que remetem para referências de outros registos culinários. Toda esta informação foi manualmente coligida pelo editor e registada neste campo, em cada receita, como já atrás referido.
- **Procedimento** — Que representa a descrição de todo o processo de elaboração da receita.
- **Processos** — Que são o resultado do processo de extração, em forma de lista. Na prática, é um conjunto de atos de culinária que estão direta ou indiretamente identificados no texto. Aqui são apresentados os processos na sua forma canónica e não como se encontram explicitamente escritos.

Figura 11: Importação de receitas — o segundo passo.

Após o avanço no processo, o sistema revela as palavras que serão indexadas na classe Índice da ontologia, bem como o número de ocorrências denotadas na receita.

No derradeiro passo da importação (Figura 12), é apresentado um pequeno resumo do conhecimento que será introduzido na base ontológica. Relativamente aos processos recolhidos, este último passo apresenta o processo na forma de verbo no infinitivo, e a sua forma explícita no procedimento. Em qualquer momento da importação, o utilizador pode cancelar o processo e o sistema redireciona para a página inicial. Caso se confirme a opção de avançar (Figura 13), a nova receita é acrescentada. Ao fazê-lo, a interligação de dados evolui tanto ao nível do Neo4J como no ficheiro OWL.

Figura 12: Importação — Último passo — Parte I

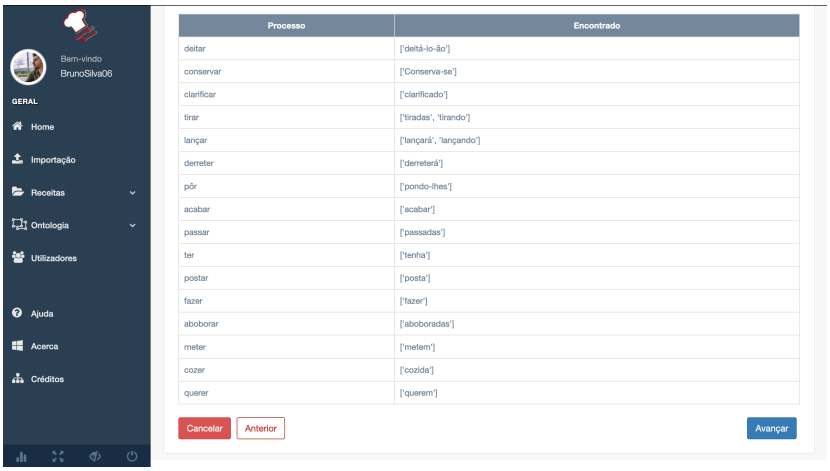

Figura 13: Importação — Último passo — Parte II

- **Pesquisa** (Figura 14) – Funcionalidade situada no menu *dropdown* Ontologias, que permite procurar receitas no modelo de dados até então desenvolvido. Nesta opção permite-se fazer a pesquisa por título, ingrediente ou processo que estejam implícitos no documento da receita, e, no caso dos processos, o processo encontrado no texto sob forma de verbo no infinitivo.

  Uma vez pressionado o botão de pesquisa, com base nos parâmetros especificados, a aplicação apresenta uma grelha de resultados, na qual cada registo apresentado corresponde a um cartão alusivo à receita. Este cartão contém o ID, a imagem do texto manuscrito (opcional), o título, a data em que foi integrado o documento culinário e uma indicação dos parâmetros que foram encontrados. Adicionalmente, esta funcionalidade permite efetuar pesquisas diretas no browser do Neo4J.

  Para consultar os detalhes de cada registo, cada cartão disponibiliza um botão que redireciona o utilizador para uma página com a informação extraída e relativa à receita em exposição.

Figura 14: Resultados de um processo de pesquisa

- **Overview** (Figura 15) — Esta funcionalidade permite visualizar o modelo conceptual orientado à ontologia, identificando univocamente os elos de ligação, bem como as classes presentes no sistema ontológico. Este modelo tem como objetivo representar a forma como as classes se ligam entre si e como estas ligações são usadas para efetuar pesquisas sobre o modelo de dados.

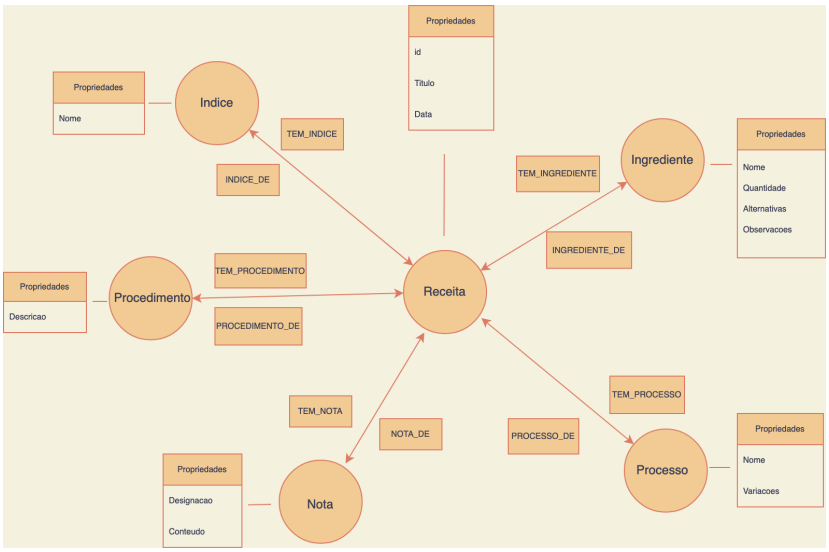

Figura 15: O modelo conceptual da ontologia

4.1.1 *Detalhes da Receita*

As características de uma receita estão organizadas, essencialmente, em quatro blocos (Figura 16):

- **Cabeçalho** — Espaço dedicado ao identificador, título da receita e data na qual a receita foi inserida.
- **Ingredientes** — Bloco, da responsabilidade do editor, no qual são apresentados os ingredientes encontrados na receita.
- **Preparação** — Um texto que descreve os momentos-chave na confeção da receita. Além disso, são ainda identificados, a vermelho, todos os processos ou ações que foram caracterizados pelo sistema de extração automática.
- **Processos** — Conjunto de processos, ou cada um dos atos culinários, que foram identificados no bloco anterior.

Figura 16: Receita — Detalhes

4.1.2 *Sistema Ontológico*

Na arte da culinária, muitas vezes surge a necessidade de consultar outras receitas que usem um determinado ingrediente ou processo, com a finalidade de perceber de que forma são usados, sobretudo porque as receitas antigas eram frequentemente omissas ou muito sucintas. Neste aspeto, o sistema ontológico permite compreender como estes processos ou ingredientes são realmente utilizados, na medida em que cada termo está diretamente relacionado com o resto do domínio, que, neste caso, se assume como uma forma de

culinária. Na prática, a aplicação possui um mecanismo de interligação de termos, que faculta ao utilizador uma experiência de utilização mais intuitiva e eficaz.

- **Ingredientes** — Cada ingrediente representa uma hiperligação com o resto das receitas que contenham o registo selecionado. Ao selecionar-se um ingrediente, o sistema apresenta uma janela com mais detalhes sobre o ingrediente da receita, possibilitando navegar por um conjunto de resultados que tenham esse termo.
- **Processos** — O conjunto de termos assinalados a vermelho são também hiperligações para um conjunto de resultados (Figura 17) que integrem o processo selecionado. Além desta funcionalidade, os elementos caracterizados no separador dos processos contêm uma hiperligação para o significado do processo. Esta última *feature* não recorre ao sistema ontológico, mas sim a um domínio externo.

Figura 17: Conjunto de resultados resultantes da hiperligação do processo "cozer"

4.1.3 *Importação de uma Receita*

De forma a perceber todo o processo de análise e extração de conhecimento por parte do sistema, vejamos como importar e analisar uma receita em concreto. Para isso, selecionámos a receita 79, “Salmão e Solho” (Figura 18). O motivo da escolha desta receita explica-se pelos diversos casos particulares que ela possui. Estando o utilizador no menu de importação, ele pode selecionar o documento que contém esta receita e, opcionalmente, pode inserir uma imagem do fólio manuscrito em que figura a receita.

De seguida, o sistema verifica se a receita já foi importada, usando o seu identificador, e caso exista, não permite a sua importação. Caso contrário, o sistema permitirá a introdução da receita, fazendo o extrator a extração do identificador (79) e do título da mesma (Salmão e Solho). Após esta extração é iniciado o processamento do bloco que corresponde aos

```
                                        79
                                    Salmão e solho

Ingredientes:
Salmão (para cozer)
Vd. Trutas:
        1/3 de vinagre
        salsa
        sal
Solho (para cozer)
Ervas aromáticas (não especificadas)
Vinagre
Cebola
Sal
Pimenta
Solho refogado, com molho:
Vd. Peixes de molho (r. 74)
Para conservar salmão e outros peixes crus:
Urtigas

Coze-se e tempera-se como as trutas, e assim este como o mais peixe se tem muito metendo-lhe urtigas
por dentro e na boca e mais aberturas, e embrulhando-o nelas. O solho se faz de molho, e também se
coze com cheiros, vinagre, cebola, sal, pimenta, e com ela se come.
```

Figura 18: Receita 79 — Salmão e Solho

ingredientes, uma vez que este pode conter informação adicional, identificada através da pontuação e escrita. Aqui, o extrator separa os ingredientes de eventuais notas existentes, de forma que ambos sejam tratados diferentemente. Estas notas são caracterizadas por uma designação, que corresponde ao texto contido entre o início da linha e o sinal de pontuação *:*, e o conteúdo da nota, que é constituída por todos os registos posteriores à designação e que estejam indentados (Tabela 2). Caso sejam seguidos por outra nota, extrai todos os registos até à próxima designação.

Por outro lado, os ingredientes podem aparecer enumerados numa única linha. Neste caso, o sistema começa por verificar se as palavras se encontram separadas por vírgulas ou pela conjunção *e*. De seguida, são identificadas possíveis observações contidas entre parênteses. Caso existam referências às quantidades de cada ingrediente, estas são extraídas, quer sejam apenas números ou conjuntos (e.g. dúzia). Relativamente aos ingredientes, por fim, é averiguada a possível existência de alternativas ao mesmo ingrediente, através da conjunção disjuntiva *ou* (Tabela 3).

Designação	**Conteúdo**
Vd. Trutas	1/3 de vinagre salsa sal
Solho refogado, com molho	Vd. Peixes de molho (r. 74)
Para conservar salmão e outros peixes crus	Urtigas

Tabela 2: Receita 79 — Notas

Relativamente ao bloco de texto que compõe a preparação da receita, é necessário fazer uma análise mais profunda, por forma a entender quais os processos que existem e como se relacionam com outros conceitos. Este bloco de texto é separado por tokens. Isto é, todos os potenciais termos são extraídos singularmente, juntamente com a sua função sintática. Este

Nome	Quantidade	Observações	Alternativas
salmão		para cozer	
solho		para cozer	
ervas aromáticas		não especificadas	
vinagre			
cebola			
sal			
pimenta			

Tabela 3: Receita 79 — Ingredientes

processo desempenha um papel fulcral na definição dos padrões de Hearst (Hearst, 1992). Estes padrões são aplicados ao texto que descreve os momentos da confeção do prato. Para além destes padrões, são comparados os termos extraídos (Tabela 4) com os processos já existentes na ontologia, o que confere uma natureza evolutiva àontologia. Aplicado o algoritmo, os termos resultantes são reduzidos à sua forma canónica para que o processo seja o mais claro e genérico possível (Tabela 5).

Processo Geral	Processo literal	Foi extraído?
Cozer	Coze-se	sim
Temperar	Tempera-se	sim
Meter	metendo-lhe	sim
Embrulhar	embrulhando-o	sim
Comer	come	sim

Tabela 4: Receita 79 — Processos

	Expressão capturada	Processo
Padrão 1	Coze-se	cozer
Padrão 1	tempera-se	temperar
Padrão 1	metendo-lhe	meter
Padrão 1	embrulhando-o	embrulhar
Padrão 3	tempera-se com	temperar
Padrão 3	coze com	cozer
Padrão 4	se e tempera-se	temperar
Padrão 7	se tem	ter
Padrão 7	se faz	fazer
Padrão 7	se coze	cozer
Padrão 7	se come	comer
Padrão 10	e embrulhando-o	embrulhar
Padrão 10	e faz	fazer
Padrão 10	e coze	cozer

Tabela 5: Receita 79 — Padrões de Hearst para obter processos culinários

A derradeira etapa de todo o processo de importação revela-se sob a forma de um pequeno resumo (Figura 19), no qual está combinada a informação extraída, nomeadamente o número de palavras indexadas e a foto do fólio manuscrito com a receita.

Figura 19: Receita 79 — Resumo da Importação I

Ao confirmar a importação desta receita no sistema, a base ontológica será atualizada com todos os conceitos e relações da receita “Salmão e Solho”. A fusão é efetuada através do comando MERGE da linguagem *Cypher*. Este comando verifica se o nodo desta receita existe na base de dados. Caso exista, atualiza apenas as relações que se reportam ao nodo em questão. Este mesmo processo é efetuado para todos os conceitos que se relacionam com a receita. Desta forma, é garantido que a inserção de novas entidades na ontologia apenas seja realizada quando estas não existam, permitindo que não haja entidades duplicadas. Uma vez concluído o processo de importação, a receita, e todas as relações inerentes (Figura 20), podem ser consultadas no menu de pesquisa.

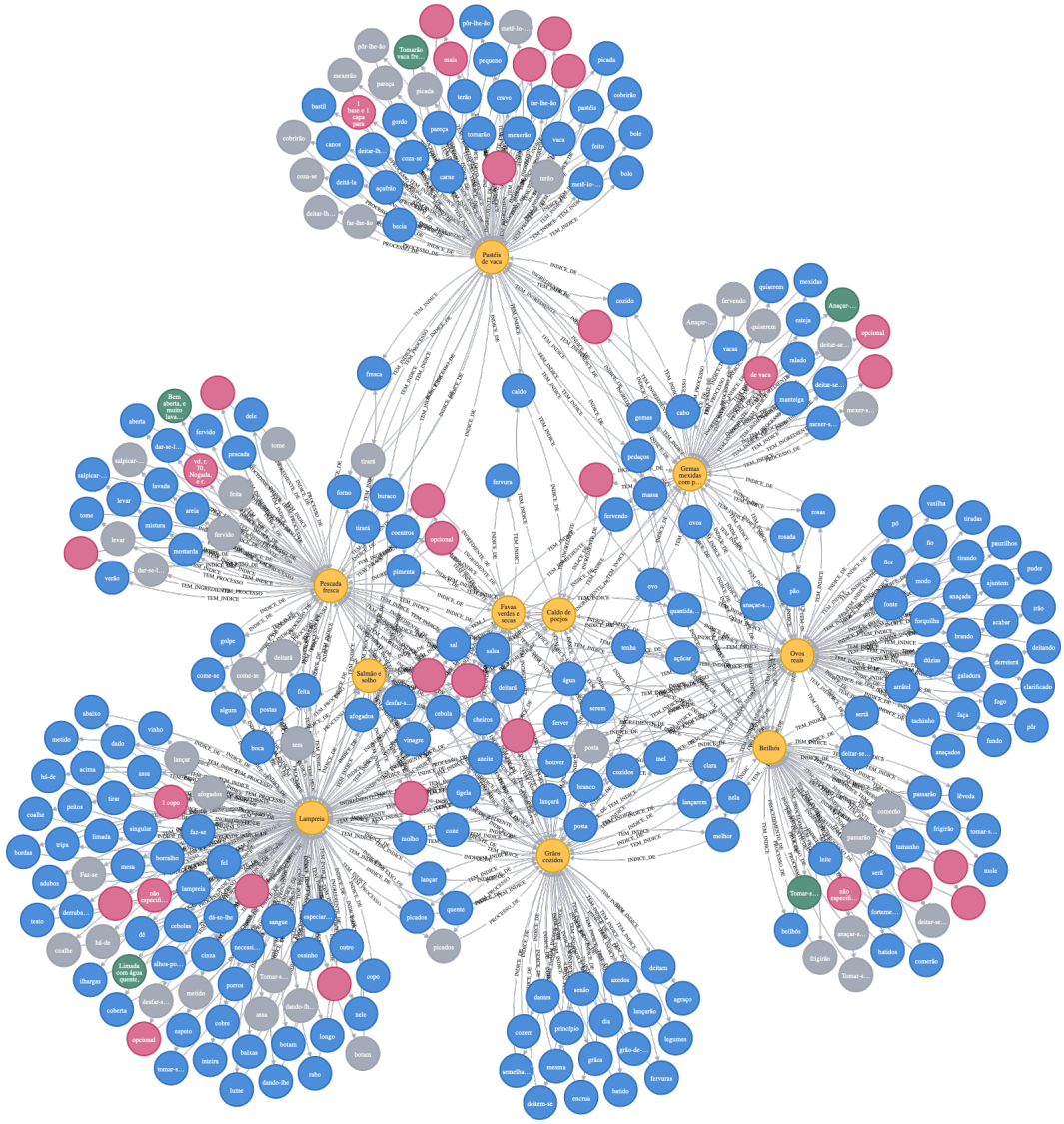

Figura 20: Receita 79 — Ontologia (Neo4j)

4.2 ANÁLISE DA ONTOLOGIA

O povoamento da ontologia evidencia as interligações que imprimem a forma como cada conceito se relaciona com os restantes no domínio da culinária. Inicialmente, o crescimento da ontologia será grande a nível de novas entidades. Quando o número de receitas importadas atingir um número considerável, o crescimento passará a ser mais notório ao nível das relações entre conceitos. Este facto explica-se pelo número finito de processos de culinária possíveis (apesar de numerosos) e pelas quase infinitas combinações de relações entre termos. Esta evolução da ontologia permite refinar o processo de extração, que, ao longo do tempo, será cada vez mais fiável e mais correto. Ainda na perspetiva qualitativa, a aplicação de algoritmos baseados em padrões léxicos e sintáticos, em vez do emprego de algoritmos estatísticos, permitiu melhorar significativamente a caracterização dos processos culinários.

Conceptualmente, este modelo ontológico (ver secção 3.3) permite uma organização mais prática dos conteúdos relativos à culinária, bem como perceber quais as entidades que representam os conceitos pertencentes a cada receita. A quantidade de ingredientes poderá revelar uma receita complexa se, simultaneamente, a quantidade de processos for igualmente grande. A existência de notas sugere uma possível ligação entre receitas, provavelmente a nível de ingredientes, processos ou simplesmente de uma palavra. Na Figura 21, é possível perceber o volume de informação interligada somente com a receita 45 — "Pastéis de Vaca".

No panorama geral, a ontologia extrai os elementos mais pertinentes para a confeção de cada receita presente no livro de cozinha/(edição do) manuscrito. O uso deste modelo de dados, neste caso de estudo em particular, facilitou a interpretação da língua, o estudo linguístico e a exploração dos costumes gastronómicos relativos à época quinhentista.

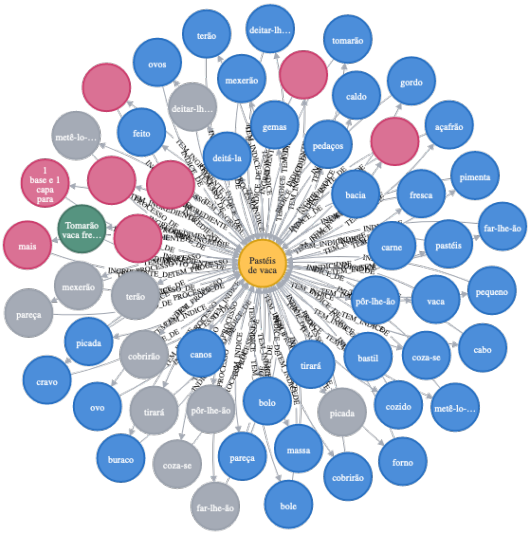

Figura 21: Receita 45 — "Pastéis de Vaca"

4.3 UTILIDADE E VIABILIDADE

A plataforma desenvolvida expõe toda a informação extraída de um forma simplificada, através de uma ontologia, e também oferece ao utilizador uma experiência muito mais frutuosa. Num cenário de aplicação real, esta plataforma facilita substancialmente o processo de pesquisa, quer de uma receita e dos seus componentes, quer de um componente e de todas as receitas ligadas a ele. Este componente pode surgir em forma de termo, ingrediente ou de um simples processo utilizado na confeção da receita. A sua utilidade estende-se ainda à identificação de processos que o próprio utilizador desconheça, mesmo após a leitura do conjunto de etapas associadas à confeção do prato. Tal acontece porque a ontologia é gerada com base num processo de extração que apenas olha para o contexto no qual os termos aparecem, colocando em segundo plano a sua interpretação literal.

Relativamente à semântica, a ontologia desenvolvida permite estabelecer uma ponte entre processos antigos e processos da era moderna que, de uma forma ou de outra, usualmente são semelhantes ou, por vezes, são os mesmos. Um raciocínio idêntico pode-se aplicar aos ingredientes, que, outrora, foram utilizados em receitas, mas que nos dias de hoje já se encontram substituídos por outros. Na prática, esta ontologia permite identificar padrões de culinária do século XVI e compará-los com as práticas de culinária atuais, algo valioso para os investigadores da área da História da Alimentação, que se dedicam a investigá-las e compará-las ao longo do tempo e entre civilizações e culturas diferentes, pelo menos desde a antiguidade greco-latina.

Relativamente a este sistema de extração automática, no decorrer das importações das receitas para a respetiva ontologia, a taxa de sucesso de extração de termos pretendidos foi elevada, mas nem sempre conseguida de forma completa. No entanto, esta taxa irá aumentando à medida que novas receitas forem sendo introduzidas no sistema. Mesmo que um termo não seja extraído pelos padrões léxicos e sintáticos, se ele já existir associado a um conceito, então será objeto de extração.

A viabilidade da ontologia tem um espectro tão largo quanto o seu domínio e poderá ser facilmente adaptada a um outro sistema. Os conceitos são transversais a qualquer documento específico do domínio da culinária e, por isso, a ontologia é igualmente transversal a qualquer caso de estudo do foro da culinária. Contudo, o mesmo não se aplica ao processo de extração. Este teria que ser revisto porque, apesar de o domínio ser o mesmo, o contexto e o estado da língua diferem, o que resulta em padrões léxicos e sintáticos diferentes. Além de viável, esta ontologia pode servir como ontologia base para outros sistemas de extração automática. Desta forma, poder-se-á utilizar o conhecimento contido na ontologia para um treino preliminar dos dados disponíveis, possibilitando abordagens que requeiram conhecimento prévio para o processo de extração, tal como acontece em abordagens estatísticas.

5

CONCLUSÕES E TRABALHO FUTURO

5.1 CONCLUSÕES

Nesta dissertação foi abordada a temática das ontologias e o uso das mesmas em sistemas de extração automática, mais concretamente em sistemas de extração automática de ontologias em textos de culinária não estruturados. O sistema desenvolvido permitiu explorar os conteúdos das receitas, editadas em lição semidiplomática (com a ortografia original) e em lição interpretativa (com grafia atualizada e alguns aspetos morfológicos adaptados à língua contemporânea) a partir de um manuscrito inédito, anónimo e sem data, na obra intitulada "As receitas de cozinha de um frade português do século XVI", e apresentá-los de uma forma mais organizada e interligada, explorando os benefícios da ontologia automaticamente gerada pelo sistema. A metodologia adotada para o desenvolvimento do sistema foi sustentada por uma análise rigorosa, que tomou em consideração as várias possíveis alternativas, entre elas a escolhida, que foi a que obteve melhores resultados em termos de precisão de termos extraídos por receita. A escolha das tecnologias usadas foi também um fator preponderante na obtenção do produto final, apesar de não termos muitas outras alternativas.

Todas as etapas do desenvolvimento do sistema desenvolvido no âmbito desta dissertação foram planeadas e executadas conforme previsto na metodologia definida, o que resultou numa rápida organização das tarefas a executar. Ao realizarmos este trabalho, foi elucidativo e notório que, na área do processamento de informação, as técnicas têm um espectro de possibilidades bastante alargado e saber escolhê-las nem sempre é tarefa fácil. O pré-processamento de dados tornou-se fundamental, devido à menor complexidade com que auxilia a manipulação dos dados, especialmente quando estes estão devidamente organizados e catalogados. A fase de extração de termos, além de ter sido a mais demorada (e mais importante) desta dissertação, requereu a priori a análise de uma série de premissas, já que cada fonte de conhecimento tinha a sua identidade, quer ela fosse de domínio ou de cariz pessoal, do autor que a compôs.

O estudo e a comparação de algoritmos meramente estatísticos com algoritmos de padrões textuais revelaram-se bastante interessantes. O primeiro apresentou uma taxa de sucesso pouco elevada neste domínio, não só porque careceu da análise de muitas receitas, mas

também porque as primeiras extrações foram vitais para as que se lhes seguiram, tal como um efeito de bola de neve, e, por consequência, a qualidade da primeira extração terá impacto na qualidade da extração das receitas seguintes. A abordagem utilizada tirou bastante proveito de padrões textuais, claros articuladores de discurso, e do papel gramatical, ou função que cada palavra desempenha numa frase. A manutenção dos padrões definidos acabou por não ser realizada, uma vez que os artigos trabalhados estão preservados e não irão sofrer alterações no decorrer do tempo. A análise sintática de textos escritos em língua portuguesa,sobretudo de períodos pretéritos, como é o caso do português clássico aqui em estudo, torna-se uma tarefa complexa e quase nunca perfeita, uma vez que existem diversas maneiras de se construir uma frase, por vezes com elevado grau de omissão de elementos ou ordem dos elementos menos comum hoje em dia. No âmbito desta dissertação, a opção de utilizar uma biblioteca direcionada para este tipo de análise, aliado ao facto de a escrita dos documentos analisados ser relativamente homogénea, facilitaram muito este processo.

Como sabemos, o desenvolvimento de ontologias continua em franca expansão. Não obstante, a construção de um sistema ontológico que seja escalável e de fácil interpretação continua a ser, também, um grande desafio. A Web semântica veio dar um grande impulso e mais visibilidade a este tipo de sistemas, devido à necessidade de tratamento de informação interligada. Já existindo esta área de trabalho e investigação há alguns anos, os recursos e ferramentas são cada vez melhores para a exploração e correlação de dados.

O desenvolvimento deste sistema de receitas permitiu evidenciar as diferenças entre a criação ontológica manual e um sistema capaz de analisar um conjunto de dados e construir uma base de conhecimento de forma (semi)automática. Ficaram evidentes as vantagens que um sistema de integração automático proporciona, não só na análise dos elementos constituintes da área de domínio, mas também na facilidade que permite ao utilizador de pesquisa sobre um determinado termo. Contudo, é necessário refletir e avaliar todas as contrapartidas de um sistema desta natureza. Construir um modelo de dados para este tipo de sistema requer regras e ideologias próprias do domínio, que servem apenas e só o propósito de integrar num sistema que opere sobre uma base de conhecimento. Aliados a este facto, estão o tempo e os recursos necessários para elaborar todo o processo de extração, que, caso seja efetuado numa fonte em evolução, necessita de ser revisto e por vezes integralmente reformulado. Adicionalmente, a extração automática de ontologias terá sempre um fator de precisão que irá variar conforme os padrões textuais ou a dimensão linguística, que se vão mantendo ou alterando. Eventualmente, será possível chegar a uma precisão quase completa, mas nunca antes de uma longa fase de análise e desenvolvimento de regras e condições aplicáveis aos textos em causa.

Em concreto, neste trabalho, o sistema desenvolvido apresenta mais vantagens do que desvantagens, uma vez que os textos permanecem inalterados há anos e a sua estrutura é relativamente homogénea em todos os documentos analisados, sobretudo na lição inter-

pretativa com o arranjo do editor (destacando os Ingredientes). Isto não só facilita a análise como sustenta a longevidade deste sistema.

5.2 TRABALHO FUTURO

Embora a prova de conceito tenha sido elaborada e comprovada, existem ainda alguns fatores que devem ser integrados, no sentido de aperfeiçoar a análise e desempenho do sistema. Uma das melhorias passa por distinguir o que são processos de culinária e ações do utilizador, devido ao facto de uns verbos identificarem processos inerentes aos ingredientes e outros corresponderem a etapas que devem ser seguidas pela pessoa que esteja a confecionar o prato. Esta distinção implicaria uma análise sintática ainda mais precisa e detalhada, uma vez que teriam que ser analisados os elementos que precedem e os que sucedem o verbo, e considerada a evolução semântica da língua, já que certas formas e construções do português clássico podem ser enganadoramente interpretadas por um utilizador do português contemporâneo, mesmo apresentando já ortografia atualizada. Tal análise, por vezes, tornar-se-ia complicada pelas diversas conotações que as palavras, na língua portuguesa, podem tomar, vulgarmente falando.

Um sistema que aplique a metodologia de *Ontology Learning Layer Cake* será tão bom quanto a qualidade de cada processo aplicado em cada degrau da sua pirâmide representativa. Algo que este sistema não contempla é a agregação de sinónimos aplicada aos termos, a qual permitiria uma associação mais correta dos termos a conceitos e, também, numa aplicação prática, pesquisar receitas através de um termo e dos seus sinónimos. Para alcançar esta agregação, seria preciso conceber mais regras para a caracterização de sinónimos, para além de uma melhor compreensão do significado literal e dos significados conotativos e contextuais de determinada palavra.

No mundo da culinária escrita, e sobretudo manuscrita, é hábito o autor do manuscrito acrescentar especificações em vários momentos, ou, posterirmente, o editor do mesmo anotar observações, ou explicar o que determinadas etapas requerem, com o intuito de providenciar ao leitor uma clarividência de todo o enredo de uma receita. Durante o estudo e análise dos textos, foi aproveitado o glossário incluído no final do livro, e que explica determinados processos ou conceitos na arte da culinária, e também referências de outras receitas, que, no entanto, não foram contemplados no sistema ontológico. Este glossário, em termo práticos, poderia ligar processos a outras receitas e não apenas redirecionar o utilizador para a definição do processo. Além disso, poderiam ser explicados atos culinários antigos ou que fossem sinónimos de ações mais conhecidas. A sua integração futura no sistema teria, pois, especial interesse, mesmo nos processos de interpretação, quer semântica quer estrutural, e de desambiguação contextual.

Na temática das ontologias, a manutenção de dois modelos ontológicos — o ficheiro OWL e o que se encontra registado em Neo4J — torna-se, a longo prazo, dispendiosa e não tão eficiente quanto poderia ser. Neste sentido, é importante que possa evoluir para um mecanismo que converta todo o conhecimento existente na base de dados em Neo4j para um formato desejado pelo utilizador. Esta funcionalidade permitiria a incorporação dos dados obtidos neste sistema em outros sistemas com características diferentes. Tal seria bastante proveitoso.

Por último, o sistema tem potencial para evoluir, não só estética e funcionalmente, mas também no sentido tecnológico, em que poderiam ser incorporadas interfaces mais reativas e atuais. Está mais do que comprovado que uma interface intuitiva e apelativa, promove uma experiência mais imersiva e, consequentemente, a continuidade da utilização do sistema por parte dos utilizadores. Além do aspeto visual, há margem para incluir preferências de utilização no sistema, através do aperfeiçoamento da atual ontologia, que iria permitir uma filtragem de resultados mais pertinente e prática.

BIBLIOGRAFIA

Ahmad, K., & Gillam, L. (2005). Automatic ontology extraction from unstructured texts. *In* Lecture Notes in Computer Science (pp. 1330–1346). Springer Berlin Heidelber

Alfonseca, E., & Manandhar, S. (2002). Extending a lexical ontology by a combination of distributional semantics signatures. Em Knowledge Engineering and Knowledge Management: Ontologies and the Semantic Web (pp. 1–7). Springer Berlin Heidelberg.

Aussenac-Gilles, N., Biébow, B., & Szulman, S. (2000). Revisiting ontology design: A method based on corpus analysis. Em Knowledge Engineering and Knowledge Management Methods, Models, and Tools (pp. 172–188). Springer Berlin Heidelberg.

Bachimont, B., Isaac, A., & Troncy, R. (2002). Semantic commitment for designing ontologies: A proposal. Em Knowledge Engineering and Knowledge Management: Ontologies and the Semantic Web (pp. 114–121). Springer Berlin Heidelberg.

Barros, A. L. (with Seiça, M., Veloso, J. & Aguiar, M.) (2013), As receitas de cozinha de um frade português do século XVI. Coimbra: Imprensa da Universidade de Coimbra.

Beckwith, R., Fellbaum, C., Gross, D., & Miller, G. A. (2021). WordNet: A lexical database organized on psycholinguistic principles. Em Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon (pp. 211–232). Psychology Press.

Biébow, B., Szulman, S., & Clément, A. J. B. (1999). TERMINAE: A linguistics-based tool for the building of a domain ontology. Em Knowledge Acquisition, Modeling and Management (pp. 49–66). Springer Berlin Heidelberg.

Buitellar, P., Cimiano, P., Frank, A., Racioppa, S. (2006). SOBA: SmartWeb Ontology-based Annotation. Em AIFB, University of Karlsuhe, Germany.

Choudhary, J., & Tomar, D. S. (2014). Semi-Automated Ontology building through Natural Language Processing. International Journal of Computers & Technology, 13(8), 4738–4746. https://doi.org/10.24297/ijct.v13i8.7072

Davies, J., Fensel, D., & van Harmelen, F. (2003). Conclusions: Ontology-driven knowledge management – towards the semantic web? Em Towards the Semantic Web (pp. 265–266). John Wiley & Sons, Ltd.

El Ghosh, M., Naja, H., Abdulrab, H., & Khalil, M. (2017). Ontology learning process as a bottom-up strategy for building domain-specific ontology from legal texts. Proceedings of the 9th International Conference on Agents and Artificial Intelligence.

Fitzpatrick, D. (sem data). Graph data platform. Neo4j Graph Data Platform; Neo4j. https://neo4j.com/

Fortuna, Blaž Grobelnik, Marko Mladenić, Dunja. (2007). OntoGen: Semi-automatic Ontology Editor. 309-318.

Glimm, B., Horrocks, I., Motik, B., Stoilos, G., Wang, Z. (2014). HermiT: An OWL 2 reasoner. Journal of Automated Reasoning, 53(3), 245–269.

Hassan, A. Z., Vallabhajosyula, M. S., & Pedersen, T. (2018). UMDuluth-CS8761 at SemEval-2018 Task9: Hypernym Discovery using Hearst Patterns, Co-occurrence frequencies and Word Embeddings. Proceedings of The 12th International Workshop on Semantic Evaluation.

Hearst, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. Proceedings of the 14th conference on Computational linguistics.

Horrocks, I. (2013). What Are Ontologies Good For? Em Evolution of Semantic Systems (pp. 175–188). Springer Berlin Heidelberg.

Khan, L. & Luo, F. (2002). Ontology construction for information selection. IEEE Transactions on Applications and Industry. 122- 127. 10.1109/TAI.2002.1180796.

Kietz, J.-U., Volz, R., & Maedche, A. (2000). Extracting a domain-specific ontology from a corporate intranet. Proceedings of the 2nd workshop on Learning language in logic and the 4th conference on Computational natural language learning.

Maedche, A. & Staab, S. (2000). The TEXT-TO-ONTO Ontology Learning Environment.

Mankovskii, S., Gogolla, M., Urban, S. D., Dietrich, S. W., Urban, S. D., Dietrich, S. W., Yang, M.-H., Dobbie, G., Ling, T. W., Halpin, T., Kemme, B., Schweikardt, N., Abelló, A., Romero, O., Jimenez-Peris, R., Stevens, R., Lord, P., Gruber, T., Leenheer, P. D., ... Bechhofer, S. (2009). OWL: Web ontology language. Em Encyclopedia of Database Systems (pp. 2008–2009). Springer US.

Meslati, D., Souici Meslati, L., Mecheri, K., Boufaida, M. (2019). Context-based interoperability of semantic web services. International Journal of Metadata, Semantics and Ontologies, 13(3), 209.

Mishra, T. S. & Jain, S. (2014). Automatic Ontology Acquisition and Learning. International Journal of Research in Engineering and Technology. 03. 38-43.

Mucheroni, M. Modesto, F. (2011). A Interoperabilidade dos Sistemas de Informação sob o Enfoque da Análise Sintática e Semântica de Dados na WEB. PontodeAcesso. 5(1), 3.

Mylopoulos, John Buitelaar, Paul Olejnik, Daniel Sintek, Michael Valle, Emanuele Castagna, Paolo Brioschi, Maurizio. (2003). OntoLT: A Protégé Plug-In for Ontology Extraction from Text.

Navigli, R., Velardi, P., Cucchiarelli, A., Neri, F. (2004). Quantitative and qualitative evaluation of the OntoLearn ontology learning system. Proceedings of the 20th international conference on Computational Linguistics — COLING '04.

Nobécourt J. (2000). A method to build formal ontologies from text. Em EKAW00, Workshop on ontologies and text.

Ontotext delivers euBusinessGraph marketplace. (2019, Outubro 29). Ontotext. https://www.ontotext.com/

Reinberger, M. & Spyns, P. (2004). Discovering knowledge in texts for the learning of DOGMA-inspired ontologies. Proceedings of the ECAI Workshop on Ontology Learning and Population.

Roitman, H., & Gal, A. (2006). OntoBuilder: Fully automatic extraction and consolidation of ontologies from web sources using sequence semantics. Em Current Trends in Database Technology – EDBT 2006 (pp. 573–576). Springer Berlin Heidelberg.

Shamsfard, M., & Abdollahzadeh Barforoush, A. (2003). The state of the art in ontology learning: a framework for comparison. The knowledge engineering review, 18(4), 293–316. https://doi.org/10.1017/s0269888903000687

Stanford Center for Biomedical Informatics Research. (sem data). Protégé. Stanford.Edu. https://protege.stanford.edu/.

re — Regular expression operations — Python 3.10.5 documentation. (sem data). Python.org. https://docs.python.org/3/library/re.html